Volume 152(1) 1 January 1995 pp 27-32

Basic Statistics for Clinicians: 1. Hypothesis Testing

[Statistics]

Guyatt, Gordon; Jaeschke, Roman; Heddle, Nancy; Cook, Deborah; Shannon, Harry; Walter, Stephen

From the departments of Clinical Epidemiology and Biostatistics, Medicine and Pathology, McMaster University, Hamilton, Ont.

Drs. Guyatt and Cook are the recipients of Career Scientist Awards from the Ontario Ministry of Health. Dr. Cook is a scholar of the St. Joseph's Hospital Foundation, Hamilton, Ont. Dr. Walter is the recipient of a National Health Scientist Award from Health Canada.

Reprint requests to: Dr. Gordon Guyatt, Rm. 2C12, McMaster University Health Sciences Centre, 1200 Main St. W, Hamilton, ON L8N 3Z5

This is the first article in a series of four, to appear in the January and February 1995 issues of CMAJ.

Outline

Graphics

Abstract

In the first of a series of four articles the authors explain the statistical concepts of hypothesis testing and p values.In many clinical trials investigators test a null hypothesis that there is no difference between a new treatment and a placebo or

between two treatments. The result of a single experiment will almost always show some difference between the experimental and the control groups. Is the difference due to chance, or is it large enough to reject the null hypothesis and conclude that there is a true difference in treatment effects? Statistical tests yield a p value: the probability that the experiment would show a difference as great or greater than that observed if the null hypothesis were true. By convention, p values of less than 0.05 are considered statistically significant, and investigators conclude that there is a real difference. However, the smaller the sample size, the greater the chance of erroneously concluding that the experimental treatment does not differ from the control -- in statistical terms, the power of the test may be inadequate. Tests of several outcomes from one set of data may lead to an erroneous conclusion that an outcome is significant if the joint probability of the outcomes is not taken into account. Hypothesis testing has limitations, which will be discussed in the next article in the series.

Clinicians are often told that they are supposed to not only read journal articles, but also understand them and make a critical assessment of their validity [1,2]. Clinicians may offer better care if they are able to appraise critically the original literature and apply the results to their prac-tice [3,4]. Criteria for assessing the strength of the methods reported in medical articles can provide clinicians with guidance in recognizing the strengths and weaknesses of clinical research [5,6]. However, such guidelines tend to make only passing reference to statistical methods or interpretation of study conclusions based on statistics.

Some authors have attempted to fill this gap [7,8,9,10,11]. This series has modest goals. We do not intend, for instance, to enable readers to identify or understand the statistical tests used to calculate a p value, but we are interested in helping them interpret the p values generated by such tests. We wish to allow readers to understand the conclusions derived from statistical procedures that they find in clinical articles. This series complements our guides to using the medical literature, which focus on study design and application of study results [12].

COMMONLY USED STATISTICAL TECHNIQUES

We chose to address only the techniques and approaches that clinicians most commonly face. To identify these, we reviewed recent contributions to three major medical journals: original, special and review articles in the New England Journal of Medicine (1991; 324: 1-352); diagnosis and treatment, review, and academia articles in the Annals of Internal Medicine (1991; 114: 345-834), and original research, current review, and clinical and community studies articles in the Canadian Medical Association Journal (1991; 144: 623-1265). Two of us (N.H. and R.J.) independently reviewed 100 articles and noted the statistical techniques

used. Discrepancies between the findings of the two reviewers were resolved by consensus.

The results of this review Table 1 are consistent with those of a similar review [13]. Although a wide variety of statistical techniques were reported, hypothesis tests, confidence intervals, p values and measures of association occurred most frequently. On the basis of this information our series will deal with hypothesis testing, estimation, measures of association, survival analysis, and regression and correlation. Examples will be drawn from the articles surveyed and others.

Concept or technique	No. of articles
p value	66
Confidence interval	43
Hypothesis testing	
Parametric method	36
Nonparametric method	25
Regression or correlation	22
Measure of association	19
Survival analysis	19
Measure of agreement	8

Table 1. Frequency of statistical concepts and techniques in 100 articles published in three medicals journals

HYPOTHESIS TESTING

When we conduct a trial of a new treatment we can assume that there is a true, underlying effect of the treatment that any single experiment can only estimate. Investigators use statistical methods to help understand the true effect from the results of one experiment. For some time the paradigm for statistical inference has been hypothesis testing. The investigator starts from what is called a "null hypothesis": the hypothesis that the statistical procedure is designed to test and,

possibly, disprove. Typically, the null hypothesis is that there is no difference between outcomes as a result of the treatments being compared. In a randomized controlled trial to compare an experimental treatment with a placebo, the null hypothesis can be stated: "The true difference in the effects of the experimental and control treatments on the outcome of interest is zero."

For instance, in a comparison of two vasodilator treatments for patients with heart failure, the proportion of patients treated with enalapril who survived was compared with the proportion of survivors among patients given a combination of hydralazine and nitrates [14]. We start with the assumption that the treatments are equally effective and stick to this position unless the data make it untenable. The null hypothesis in the vasodilator trial could be stated: "The true difference in the proportion surviving between patients treated with enalapril and those treated with hydralazine and nitrates is zero."

In the hypothesis-testing framework we ask Are the observed data consistent with this null hypothesis? The logic behind this approach is the following. Even if the true difference in effect is zero, the results observed will seldom be exactly the same; that is, there will be some difference between outcomes for the experimental and control groups. As the results diverge farther and farther from the finding of no difference, the null hypothesis that there is no difference between treatments becomes less and less credible. If the difference between results in the treatment and control groups becomes large enough, the investigator must abandon belief in the null hypothesis. An explanation of the role of chance helps demonstrate this underlying logic.

THE ROLE OF CHANCE

Imagine a fair or "unbiased" coin in which the true probability of obtaining heads in any single coin toss is 0.5. If we tossed such a coin 10 times we would be surprised if we saw exactly five heads and five tails. Occasionally, we would get results very divergent from the five-to-five split, such as eight to two, or even nine to one. Very infrequently 10 coin tosses would result in 10 consecutive heads or tails.

Chance is responsible for this variation in results. Games of chance illustrate the way chance operates. On occasion, the roll of two unbiased dice (with an equal probability of rolling any number between one and six) will yield two ones, or two sixes. The dealer in a poker game will, on occasion (and much to the delight of the recipient), dispense a hand consisting of five cards of a single suit. Even less frequently, the five cards will not only belong to a single suit but will also be consecutive.

Chance is not restricted to the world of coin tosses, dice and card games. If a

sample of patients is selected from a community, chance may result in unusual distributions of disease in the sample. Chance may be responsible for a substantial imbalance in the rates of a particular event in two groups of patients given different treatments that are, in fact, equally effective. Statistical inquiry is geared to determining whether unbalanced distributions can be attributed to chance or whether they should be attributed to another cause (treatment effects, for example). As we will demonstrate, the conclusions that may be drawn from statistical inquiry are largely determined by the sample size of the study.

THE P VALUE

One way that an investigator can go wrong is to conclude that there is a difference in outcomes between a treatment and a control group when, in fact, no such difference exists. In statistical terminology, erroneously concluding that there is a difference is called a Type I error, and the probability of making such an error is designated alpha. Imagine a situation in which we are uncertain whether a coin is biased. That is, we suspect (but do not know for sure) that a coin toss is more likely to result in heads than tails. We could construct a null hypothesis that the true proportions of heads and tails are equal. That is, the probability of any given toss landing heads is 0.5, and so is the probability of any given toss landing tails. We could test this hypothesis in an experiment in which the coin is tossed a number of times. Statistical analysis of the results would address whether the results observed were consistent with chance.

Let us conduct a thought experiment in which the suspect coin is tossed 10 times, and on all 10 occasions the result is heads. How likely is this result if the coin is unbiased? Most people would conclude that this extreme result is highly unlikely to be explained by chance. They would therefore reject the null hypothesis and conclude that the coin is biased. Statistical methods allow us to be more precise and state just how unlikely it is that the result occurred simply by chance if the null hypothesis is true. The probability of 10 consecutive heads can be found by multiplying the probability of a single head (0.5) by itself 10 times: 0.5 x 0.5 x 0.5 and so on. Therefore, the probability is slightly less than one in 1000. In an article we would likely see this probability expressed as a p value: $p < 0.001$. What is the precise meaning of this p value? If the null hypothesis were true (that is, the coin was unbiased) and we were to repeat the experiment of the 10 coin tosses many times, 10 consecutive heads would be expected to occur by chance less than once in 1000 times. The probability of obtaining either 10 heads or 10 tails is approximately 0.002, or two in 1000.

In the framework of hypothesis testing the experiment would not be over, for we have yet to make a decision. Are we willing to reject the null hypothesis and conclude that the coin is biased? How unlikely would an outcome have to be before we were willing to dismiss the possibility that the coin was unbiased? In

other words, what chance of making a Type I error are we willing to accept? This reasoning implies that there is a threshold probability that marks a boundary; on one side of the boundary we are unwilling to reject the null hypothesis, but on the other we conclude that chance is no longer a plausible explanation for the result. To return to the example of 10 consecutive heads, most people would be ready to reject the null hypothesis when the observed results would be expected to occur by chance less than once in 1000 times.

Let us repeat the thought experiment with a new coin. This time we obtain nine tails and one head. Once again, it is unlikely that the result is due to chance alone. This time the p value is 0.02. That is, if the null hypothesis were true and the coin were unbiased, the results observed, or more extreme than those observed, (10 heads or 10 tails, 9 heads and 1 tail or 9 tails and 1 head) would be expected to occur by chance twice in 100 repetitions of the experiment.

Given this result, are we willing to reject the null hypothesis? The decision is arbitrary and a matter of judgement. However, by statistical convention, the boundary or threshold that separates the plausible and the implausible is five times in 100 ($p = 0.05$). This boundary is dignified by long tradition, although other choices of a boundary value could be equally reasonable. The results that fall beyond this boundary (i.e., $p < 0.05$) are considered "statistically significant." Statistical significance, therefore, means that a result is "sufficiently unlikely to be due to chance that we are ready to reject the null hypothesis."

Let us repeat our experiment twice more with a new coin. On the first repetition eight heads and two tails are obtained. The p value associated with such a split tells us that, if the coin were unbiased, a result as extreme as eight to two (or two to eight), or more extreme, would occur by chance 11 times in 100 ($p = 0.11$). This result has crossed the conventional boundary between the plausible and implausible. If we accept the convention, the results are not statistically significant, and the null hypothesis is not rejected.

On our final repetition of the experiment seven tails and three heads are obtained. Experience tells us that such a result, although it is not the most common, would not be unusual even if the coin were unbiased. The p value confirms our intuition: results as extreme as this split would occur under the null hypothesis 34 times in 100 ($p = 0.34$). Again, the null hypothesis is not rejected.

Although medical research is not concerned with determining whether coins are unbiased, the reasoning behind the p values reported in articles is identical. When two treatments are being compared, how likely is it that the observed difference is due to chance alone? If we accept the conventional boundary or threshold ($p < 0.05$), we will reject the null hypothesis and conclude that the treatment has some effect when the answer to this question is that repetitions of the experiment would

yield differences as extreme as those we have observed less than 5% of the time.

In the randomized trial mentioned earlier, treatment with enalapril was compared with treatment by a combination of hydralazine and nitrates in 804 male patients with heart failure. This trial illustrates hypothesis testing when there is a dichotomous (Yes-No) outcome, in this case, life or death [14]. During the follow-up period, which ranged from 6 months to 5.7 years, 132 (33%) of the 403 patients assigned to the enalapril group died, as did 153 (38%) of the 401 assigned to the hydralazine and nitrates group. Application of a statistical test that compares proportions (the chi squared (chi^2) test) shows that if there were actually no difference in mortality between the two groups, differences as large as or larger than those actually observed would be expected 11 times in 100 (chi squared (chi^2) = 0.11). We use the hypothesis-testing framework and the conventional cut-off point of 0.05, and we conclude that we cannot reject the null hypothesis -- the difference observed is compatible with chance.

RISK OF A FALSE-NEGATIVE RESULT

A clinician might comment on the results of the comparison of enalapril with hydralazine and nitrates as follows: "Although I accept the 0.05 threshold and therefore agree that we cannot reject the null hypothesis, I still suspect that treatment with enalapril results in a lower mortality rate than treatment with the combination of hydralazine and nitrates. The experiment leaves me in a state of uncertainty." This clinician recognizes a second type of error that an investigator can make: falsely concluding that an effective treatment is useless. A Type II error occurs when we erroneously fail to reject the null hypothesis (and, therefore, we dismiss a useful treatment).

In the comparison of treatment with enalapril and with hydralazine and nitrates, the possibility of erroneously concluding that there is no difference between the treatments looms large. The investigators found that 5% fewer patients receiving enalapril died than those receiving the alternative vasodilator regimen. If the true difference in mortality really were 5%, we would readily conclude that patients benefit from enalapril. Despite this result, however, we were unable to reject the null hypothesis.

Why were the investigators unable to conclude that enalapril is superior to hydralazine and nitrates despite having observed an important difference between the mortality rates? The study did not enrol enough patients for the investigators to be confident that the difference they observed was real. The likelihood of missing an important difference (and making a Type II error) decreases as the sample gets larger. When there is a high risk of making a Type II error, we say the study has inadequate power. The larger the sample, the lower the risk of Type II error and the greater the power. Although 804 patients were recruited by the investigators

conducting the vasodilator trial, for dichotomous outcomes such as life or death very large samples are often required to detect small differences in the effects of treatment. For example, the trials that established the optimal treatment of acute myocardial infarction with acetylsalicylic acid and thrombolytic agents recruited thousands of patients to ensure adequate power.

When a trial fails to reject the null hypothesis ($p > 0.05$) the investigators may have missed a true treatment effect, and we should consider whether the power of the trial was adequate. In such "negative" studies, the stronger the trend in favour of the experimental treatment, the more likely the trial missed a true treatment effect [15]. We will explain more about deciding whether a trial had adequate power in the next article in this series.

Some studies are designed to determine not whether a new treatment is better than the current one but whether a treatment that is less expensive, easier to administer or less toxic yields the same treatment effect as standard therapy. In such studies (often called "equivalence studies" [16]) recruitment of an adequate sample to ensure that small but important treatment effects will not be missed is even more important. If the sample size in an equivalence study is inadequate, the investigator risks concluding that the treatments are equivalent when, in fact, patients given standard therapy derive important benefits in comparison with those given the easier, cheaper or less toxic alternative.

CONTINUOUS MEASURES OF OUTCOME

All of our examples so far have used outcomes such as Yes or No, heads or tails, or dying or not dying, that can be expressed as proportions. Often, investigators compare the effects of two or more treatments using numeric or ordinal variables such as spirometric measurement, cardiac output, creatinine clearance or score on a quality-of-life questionnaire. These outcomes are continuous: a large number of values are possible.

For example, in the study of enalapril versus hydralazine and nitrates in the treatment of heart failure the investigators compared the effect of the two regimens on exercise capacity (a continuous variable). In contrast with the effect on mortality, which showed better results with enalapril treatment, exercise capacity improved with hydralazine and nitrates but not with enalapril. The investigators compared the change in exercise capacity from baseline to 6 months in the two treatment groups with the use of a statistical test for continuous variables (Student's t-test). Exercise capacity in the group receiving hydralazine and nitrates improved more than it did in the other group, and the difference between the two groups was unlikely to have occurred by chance ($p = 0.02$). P values for Students' t-test and others like it are obtained from standard tables.

BASELINE DIFFERENCES

Authors of articles often state that hypothesis tests have been "adjusted" for baseline differences in the groups studied. Random assignment, in which chance alone dictates to which group a patient is allocated, generally produces comparable groups. However, if the investigator is unlucky, factors that determine outcome might be unequally distributed between the two groups. For example, in a trial to compare two treatments, let us say that it is known that older patients have a poorer outcome. After random assignment, the investigator discovers that a larger proportion of the older patients are assigned to one of the two treatments. This age imbalance could threaten the validity of an analysis that does not take age into account. So the investigator performs an adjustment in the statistical test to yield a p value corrected for differences in the age distribution of the two groups. In this example, readers are presented with the probability that would have been generated if the age distribution in the two groups had been the same. In general, adjustments can be made for several variables at once, and the p value can be interpreted in the regular way.

MULTIPLE TESTS

University students have long been popular subjects for experiments. In keeping with this tradition, we have chosen medical students as the subjects for our next thought experiment.

Picture a medical school in which an introductory course on medical statistics is taught by two instructors, one of whom is more popular than the other. The dean of the medical school has no substitute for the less popular faculty member. She has a particular passion for fairness and decides that she will deal with the situation by assigning the 200 first-year medical students to one instructor or the other by random assignment, in which each student has an equal chance (0.5) of being allocated to one of the two instructors.

The instructors decide to use this decision to illustrate some important principles of medical statistics. They therefore ask Do any characteristics of the two groups of students differ beyond a level that could be explained by chance? The characteristics they choose are sex, eye colour, height, grade-point average in the previous year of university, socioeconomic status and favourite type of music. The instructors formulate null hypotheses for each of their tests. For instance, the null hypothesis associated with sex distribution is as follows: the students are drawn from the same group of people; therefore, the true proportion of women in the two groups is identical. Since the investigators know in advance that the null hypothesis in each case is true, any time the hypothesis is rejected represents a false-positive result.

The instructors survey their students to determine their status on each of the six variables of interest. For five of these variables they find that the distributions are similar in the two groups, and the p values associated with statistical tests of the differences between groups are all greater than 0.10. They find that for eye colour, however, 25 of 100 students in one group have blue eyes and 38 of 100 in the other group have blue eyes. A statistical analysis reveals that if the null hypothesis were true (which it is) then such a difference in the proportion of people with blue eyes in the two groups would occur slightly less than five times in 100 repetitions of the experiment. If the investigators used the conventional boundary the null hypothesis would be rejected.

How likely is it that, in six independent hypothesis tests on two similar groups of students, at least one test would have crossed the threshold of 0.05 by chance alone? ("Independent" means that the result of a test of one hypothesis does not, in any way, depend on the results of tests of any of the other hypotheses.) This probability is calculated as follows: the probability that we would not cross the 0.5 threshold in testing a single hypothesis is 0.95; in testing two hypotheses the probability that neither one would cross the threshold is 0.95 multiplied by 0.95 (the square of 0.95); in testing six hypotheses, the probability that not a single one would cross the 0.5 threshold is 0.95 to the sixth power, or 0.74. Therefore, when six independent hypotheses are tested the probability that at least one result is statistically significant is 0.265 or approximately 1 in 4, not 1 in 20. If we wish to maintain our overall boundary for statistical significance at 0.05, we have to divide the threshold p value by six, so that each of the six tests uses a boundary value of $p = 0.008$. That is, you would reject the null hypothesis that none of the characteristics differed significantly only if any one of the differences was significant at $p < 0.008$.

There are two messages here. First, rare findings happen on occasion by chance. Even with a single test, a finding with a p value of 0.01 will happen 1% of the time. Second, we should beware of multiple hypothesis testing, because it may yield misleading results. Examples of this phenomenon abound in the clinical literature. Pocock, Hughes and Lee, [2] in a survey of 45 trials from three leading medical journals, found that the median number of endpoints was 6, and most results were tested for statistical significance. A specific example of the dangers of using multiple endpoints is found in a randomized trial of the effect of rehabilitation after myocardial infarction on quality of life [17]. The investigators randomly assigned patients to standard care, an exercise program or a counselling program and obtained patient reports on work, leisure, sexual activity, satisfaction with outcome, compliance with advice, quality of leisure and work, psychiatric symptoms, cardiac symptoms and general health. For almost all of these variables, there was no difference between the three groups. However, the patients were more satisfied with exercise than with the other two regimens, the families in the counselling group tried to protect the patients less than those in the other groups,

and work hours and frequency of sexual activity were greater at 18 months' follow-up in the counselling group than in the other groups. Does this mean that the exercise and counselling programs should be implemented because of the small number of outcomes in their favour, or that they should be rejected because most of the outcomes showed no difference? The authors concluded that their results did not support the effectiveness of either exercise or counselling programs in improving quality of life. However, a program advocate might argue that, even if only a few of the results favoured such programs, they are worth while. Hence, the use of multiple variables opens the door to controversy.

There are several statistical strategies for dealing with multiple hypothesis testing of the same data. We have illustrated one of these in a previous example: dividing the p value by the number of tests. We can also specify, before the study is undertaken, a single primary outcome on which the main conclusions will hinge. A third approach is to derive a global test statistic that combines the multiple outcomes in a single measure. Full discussion of these strategies for dealing with multiple outcomes is beyond the scope of this article but is available elsewhere [18].

LIMITATIONS OF HYPOTHESIS TESTING

Some readers may, at this point, have questions that leave them uneasy. Why use a single cut-off point when the choice of such a point is arbitrary? Why make the question of whether a treatment is effective a dichotomy (a Yes-No decision) when it may be more appropriate to view it as a continuum (from Very unlikely to be effective to Almost certain to be effective)?

We are extremely sympathetic to such readers; they are on the right track. We will deal further with the limitations of hypothesis testing in the next article, which will present an alternative approach to testing for the presence of a treatment effect and to estimating a range of plausible values of such an effect.

CONCLUSION

We avoided listing the statistical procedures used to test the null hypotheses in the studies we have cited; we do not expect readers to recognize the many methods available or to question whether the appropriate test has been chosen. Rather, we have provided a guide to interpreting p values and a warning about their interpretation when multiple outcome measures are examined. We have alluded to the limitations of hypothesis testing and the resulting p values. In the next article, which will deal with confidence intervals, we will describe complementary techniques to address some of these deficiencies.

REFERENCES

1. Department of Clinical Epidemiology and Biostatistics, McMaster University Health Sciences Centre: How to read clinical journals: I. Why to read them and how to start reading them critically. Can Med Assoc J 1981; 124: 555-558 [Context Link]

2. Pocock SJ, Hughes MD, Lee RJ: Statistical problems in the reporting of clinical trials. A survey of three medical journals. N Engl J Med 1987; 317: 426-432 [Medline Link] [BIOSIS Previews Link] [Context Link]

3. Evidence-Based Medicine Working Group: Evidence-based medicine: a new approach to teaching the practice of medicine. JAMA 1992; 268: 2420-2425 [Context Link]

4. Guyatt GH, Rennie D: Users' guides to reading the medical literature. (editorial) JAMA 1993; 270: 2096-2097 [Context Link]

5. Sackett DL, Haynes RB, Guyatt GH et al: Clinical Epidemiology, a Basic Science for Clinical Medicine, Little, Brown and Company, Boston, 1991 [Context Link]

6. Wasson JH, Sox HC, Neff RK et al: Clinical prediction rules. Applications and methodological standards. N Engl J Med 1985; 313: 793-799 [Medline Link] [BIOSIS Previews Link] [Context Link]

7. Clegg F: Introduction to statistics. I: Descriptive statistics. Br J Hosp Med 1987; 37: 356-357 [Medline Link] [Context Link]

8. O'Brien PC, Shampo MA: Statistics series. Statistical considerations for performing multiple tests in a single experiment. 1. Introduction. Mayo Clin Proc 1988; 63: 813-815 [Medline Link] [BIOSIS Previews Link] [Context Link]

9. Altman DG, Gore SM, Gardner MJ et al: Statistical guidelines for contributors to medical journals. BMJ 1983; 286: 1489-1493 [Context Link]

10. Gardner MJ, Altman DG: Estimating with confidence. BMJ 1988; 296: 1210-1211 [Context Link]

11. Gardner MJ, Altman DG: Statistics with Confidence: Confidence Intervals and Statistical Guidelines, British Medical Journal, London, England, 1989 [Context Link]

12. Oxman AD, Sackett DL, Guyatt GH for the Evidence-Based Medicine Working Group: A users' guide to the medical literature. Why and how to get started. JAMA 1993; 270: 2093-2095 [Context Link]

13. Emerson JD, Colditz GA: Use of statistical analysis in the New England Journal of Medicine. N Engl J Med 1983; 309: 709-713 [Medline Link] [Context Link]

14. Cohn JN, Johnson G, Ziesche S et al: A comparison of enalapril with hydralazine-isosorbide dinitrate in the treatment of chronic congestive heart failure. N Engl J Med 1991; 325: 303-310 [Medline Link] [BIOSIS Previews Link] [Context Link]

15. Detsky AS, Sackett DL: When was a "negative" trial big enough? How many patients you needed depends on what you found. Arch Intern Med 1985; 145: 709-715 [Medline Link] [BIOSIS Previews Link] [Context Link]

16. Kirshner B: Methodological standards for assessing therapeutic equivalence. J Clin Epidemiol 1991; 44: 839-849 [Medline Link] [BIOSIS Previews Link] [Context Link]

17. Mayou R, MacMahon D, Sleight P et al: Early rehabilitation after myocardial infarction. Lancet

1981; 2: 1399-1401 [Medline Link] [BIOSIS Previews Link] [Context Link]

18. Pocock SJ, Geller NL, Tsiatis AA: The analysis of multiple endpoints in clinical trials. Biometrics 1987; 43: 487-498 [Medline Link] [BIOSIS Previews Link] [Context Link]

Accession Number: 00002792-199501010-00016

Volume 152(2) 15 January 1995 pp 169-173

Basic Statistics for Clinicians: 2. Interpreting Study Results: Confidence Intervals

[Statistics]

Guyatt, Gordon; Jaeschke, Roman; Heddle, Nancy; Cook, Deborah; Shannon, Harry; Walter, Stephen

From the departments of Clinical Epidemiology and Biostatistics, Medicine and Pathology, McMaster University, Hamilton, Ont.

Dr. Cook is a recipient of a Career Scientist Award from the Ontario Ministry of Health. Dr. Walter is the recipient of a National Health Scientist Award from Health Canada.

Reprint requests to: Dr. Gordon Guyatt, Rm. 2C12, McMaster University Health Sciences Centre, 1200 Main St. W, Hamilton ON L8N 3Z5.

This is the second article in a series of four, to appear in the January and February 1995 issues of CMAJ.

Outline

Graphics

Abstract

In the second of four articles, the authors discuss the "estimation" approach to interpreting study results.Whereas, in hypothesis testing, study results lead the reader to reject or accept a null hypothesis, in estimation the reader can assess whether a result is strong or weak, definitive or not. A confidence interval, based on the observed result and the size of the sample, is calculated. It provides a range

of probabilities within which the true probability would lie 95% or 90% of the time, depending on the precision desired. It also provides a way of determining whether the sample is large enough to make the trial definitive. If the lower boundary of a confidence interval is above the threshold considered clinically significant, then the trial is positive and definitive; if the lower boundary is somewhat below the threshold, the trial is positive, but studies with larger samples are needed. Similarly, if the upper boundary of a confidence interval is below the threshold considered significant, the trial is negative and definitive. However, a negative result with a confidence interval that crosses the threshold means that trials with larger samples are needed to make a definitive determination of clinical importance.

In our first article in this series we explained hypothesis testing, which involves estimating the likelihood that observed results of an experiment would have occurred by chance if a null hypothesis -- that there was no difference between the effects of a treatment and a control condition -- were true. The limitations of hypothesis testing have been increasingly recognized, and an alternative approach, called estimation, is becoming more popular. Several authors [1,2,3,4,5] have outlined the concepts that we will introduce in this article, and their discussions may be read to supplement our explanation.

An example from our first article illustrates the limitations of the hypothesis-testing approach. In the results of this trial, the decision to reject the null hypothesis rests on the analysis one prefers.

INTERPRETING STUDY RESULTS: HOW SHOULD WE TREAT HEART FAILURE?

In a double-blind randomized trial, treatment with enalapril was compared with therapy with a combination of hydralazine and nitrates in 804 men with congestive heart failure [6]. During the period patients were followed up, from 6 months to 5.7 years, 33% (132/403) of the patients assigned to enalapril died, as did 38% (153/401) of those assigned to hydralazine and nitrates. The p value associated with the difference in mortality, determined by a chi squared (chi^2) test, was 0.11.

If one considered this study an exercise in hypothesis testing and adopted the usual threshold for Type I error of $p = 0.05$, one would conclude that chance is an adequate explanation for the study results. One would classify this as a "negative" study, i.e., a study showing no important difference between treatment and control groups. However, the investigators also used their data to conduct a "survival analysis," which is generally more sensitive than a test of the difference in proportions. The p value for mortality obtained from the survival analysis was

0.08, a result that leads to the same conclusion as the simpler chi squared (chi^2) test. However, the authors also reported that the p value associated with differences in mortality after 2 years ("a point predetermined to be a major end point of the trial") was 0.016.

The reader could be excused for experiencing a little confusion. Do these results mean that this is a "positive" study supporting the use of an angiotensin-converting-enzyme (ACE) inhibitor (enalapril) rather than the combination of hydralazine and nitrates or a "negative" study leaving open the choice of drug treatments?

SOLVING THE PROBLEM: CONFIDENCE INTERVALS

How can the limitations of hypothesis testing be remedied and the confusion resolved? The solution is found in an alternative approach that does not determine the compatibility of the results with the null hypothesis. This approach poses two questions: What is the single value most likely to represent the true difference between the treatment and control groups? and, given the observed difference between treatment and control groups, What is the plausible range of differences within which the true difference may lie? The second question can be answered with the use of confidence intervals. Before applying confidence intervals to resolve the issue of the benefits of enalapril versus those of hydralazine and nitrates, we will illustrate the use of confidence intervals with a coin-toss experiment similar to the one we conducted in the first article.

Suppose that we have a coin that may or may not be biased. That is, the true probability of heads on any toss of the coin may be 0.5, but it may also be as high as 1.0 in favour of heads (every toss will yield heads) or in favour of tails (every toss will yield tails). We conduct an experiment to determine the true nature of the coin.

We begin by tossing the coin twice, and we observe one head and one tail. At this point, our best estimate of the probability of heads on any given coin toss is the value we have obtained (known as the "point estimate"), which is 0.5 in this case. But what is the plausible range within which the true probability of finding a head on any individual coin toss may lie? This range is very wide, and on the basis of this experiment most people would think that the probability may be as high or higher than 0.9, or as low or lower than 0.1. In other words, if the true probability of heads on any given coin toss is 0.9, it would not be surprising if, in any sample of two coin tosses, one were heads and one tails. So, after two coin tosses we are not much further ahead in determining the true nature of the coin.

We proceed with another eight coin tosses; after a total of 10, we have observed five heads and five tails. Our best estimate of the true probability of heads on any

given coin toss remains 0.5, the point estimate. The range within which the true probability of heads may plausibly lie has, however, narrowed. It is no longer plausible that the true probability of heads is as great as 0.9; with such a high probability, it would be very unlikely that one would observe 5 tails in a sample of 10 coin tosses. People's sense of the range of plausible probabilities may differ, but most would agree that a probability greater than 0.8 or less than 0.2 is very unlikely.

On the basis of 10 coin tosses, it is clear that values between 0.2 and 0.8 are not all equally plausible. The most likely value of the true probability is the point estimate, 0.5, but probabilities close to that point estimate (0.4 or 0.6, for instance) are also likely. The further the value from the point estimate, the less likely it represents the truth.

Ten tosses have still left us with considerable uncertainty about our coin, and so we conduct another 40 repetitions. After 50 coin tosses, we have observed 25 heads and 25 tails, and our point estimate remains 0.5. We now believe that the coin is very unlikely to be extremely biased, and our estimate of the range of probabilities that is reasonably consistent with 25 heads in 50 coin tosses is 0.35 to 0.65. This is still a wide range, and we may persist with another 50 repetitions. If after 100 tosses we had observed 50 heads we might guess that the true probability is unlikely to be more extreme than 0.40 or 0.60. If we were willing to endure the tedium of 1000 coin tosses, and we observed 500 heads, we would be very confident (but still not certain) that our coin is minimally, if at all, biased.

In this experiment we have used common sense to generate confidence intervals around an observed proportion (0.5). In each case, the confidence interval represents the range within which the truth plausibly lies. The smaller the sample, the wider the confidence interval. As the sample becomes larger, we are increasingly certain that the truth is not far from the point estimate we have observed from our experiment.

Since people's "common-sense" estimate of the plausible range differs considerably, we can turn to statistical techniques for precise estimation of confidence intervals. To use these techniques we must be more specific about what we mean by "plausible." In our coin toss example we could ask What is the range of probabilities within which, 95% of the time, the true probability would lie? The actual 95% confidence intervals around the observed proportion of 0.5 for our coin toss experiment are given in Table 1. If we do not need to be so certain, we could ask about the range within which the true value would lie 90% of the time. This 90% confidence interval, also presented in Table 1, is somewhat narrower.

Number of coin tosses	Observed result	95% confidence interval	90% confidence interval
2	1 head, 1 tail	0.01–0.99	0.03–0.98
10	5 heads, 5 tails	0.19–0.81	0.22–0.78
50	25 heads, 25 tails	0.36–0.65	0.38–0.62
100	50 heads, 50 tails	0.40–0.60	0.41–0.59
1000	500 heads, 500 tails	0.47–0.53	0.47–0.53

Table 1. Confidence intervals around a proportion of 0.5 in a coin-toss experiment

The coin toss example also illustrates how the confidence interval tells us whether the sample is large enough to answer the research question. If you wanted to be reasonably sure that any bias in the coin is no greater than 10% (that is, the confidence interval is within 10% of the point estimate) you would need approximately 100 coin tosses. If you needed greater precision -- with 3% of the point estimate -- 1000 coin tosses would be required. To obtain greater precision all you must do is make more measurements. In clinical research, this involves enrolling more subjects or increasing the number of measurements in each subject enrolled. (But take care: increasing precision by enlarging the sample or increasing the number of measurements does not compensate for poor study design [7,8,9].)

USING CONFIDENCE INTERVALS TO INTERPRET STUDY RESULTS

How can confidence intervals help us interpret the results of the trial to determine different effects of vasodilators in the treatment of heart failure? In the ACE-inhibitor arm of the trial 33% of the patients died, and in the group assigned to hydralazine and nitrates 38% died, yielding an absolute difference of 5%. This difference is the point estimate, our best single estimate of the benefit in lives saved from the use of an ACE inhibitor. The 95% confidence interval around this difference is -1.2% to 12%.

How can we now interpret the study results? The most likely value for the difference in mortality between the two vasodilator regimens is 5%, but the true difference may be up to 1.2% in favour of hydralazine and nitrates or up to 12% in favour of the ACE inhibitor. Values farther from 5% are less and less probable. We can conclude that patients offered ACE inhibitors most likely (but not certainly) will die later than patients offered hydralazine and nitrates; however, the magnitude of the difference in expected survival may be trivial or large. This way

of understanding the results avoids the Yes-No dichotomy that results from hypothesis testing, the expenditure of time and energy to evaluate the legitimacy of the authors' end point of mortality after 2 years, and consideration of whether the study is "positive" or "negative" on the basis of the results. One can conclude that, all else being equal, an ACE inhibitor is the appropriate choice for patients with heart failure, but that the strength of this inference is weak. The toxic effects and cost of the drugs, and evidence from other studies, would all bear on the treatment decision. Since several large randomized trials have now shown that a benefit is gained from the use of ACE inhibitors in patients with heart failure, [10,11] one can confidently recommend this class of agents as the treatment of choice.

INTERPRETING TRIALS THAT APPEAR TO BE "NEGATIVE"

In another example of the use of confidence intervals in interpreting study results, Sackett and associates [12] examined results from the Swedish Co-operative Stroke Study, a trial designed to determine whether patients with cerebral infarcts would have fewer subsequent strokes if they took acetylsalicylic acid (ASA) [13]. The investigators gave placebos to 252 patients in the control group, of whom 7% (18) had a subsequent nonfatal stroke, and ASA to 253 patients in the experimental group, of whom 9% (23) had a nonfatal stroke. The point estimate was therefore a 2% increase in strokes with ASA prophylaxis. The results certainly did not favour the use of ASA for prevention of stroke.

The results of this large trial, involving more than 500 patients, may appear to exclude any possible benefit from ASA. However, the 95% confidence interval around the point estimate of 2% in favour of placebo is from 7% in favour of placebo to 3% in favour of ASA. If, in fact, 3% of patients who had strokes would have been spared if they had taken ASA, one would certainly want to administer the drug. By treating 33 patients, one stroke could be prevented. Thus, one can conclude that the Swedish study did not exclude a clinically important benefit and, in that sense, did not have a large enough sample.

As this example emphasizes, many subjects are needed in order to generate precise estimates of treatment effects; this is why clinicians are turning more and more to rigorous meta-analyses that pool data from the most valid studies [14]. In the case of ASA prophylaxis for recurrent stroke, such a meta-analysis showed that antiplatelet agents given to patients with a previous transient ischemic attack (TIA) or stroke reduced the risk of a subsequent TIA or stroke by approximately 25% (confidence interval approximately 19% to 31%). This benefit is great enough that most clinicians will want to treat such patients with ASA [15].

This example also illustrates that, when one sees results of an apparently "negative" trial (one that, in a hypothesis-testing framework, would fail to exclude

the null hypothesis), one should pay particular attention to the upper end of the confidence interval, that is, the end that suggests the largest benefit from treatment. If even the smallest benefit of clinical importance lies above the upper boundary of the confidence interval, the trial is definitively negative. In contrast, if clinically important benefits fall within the confidence interval, the trial has not ruled out the possibility that the treatment is worth while.

INTERPRETING TRIALS THAT APPEAR TO BE "POSITIVE"

How can confidence intervals provide information about the results of a "positive" trial -- results that, in the previous hypothesis-testing framework, would be definitive enough to exclude chance as the explanation for differences between results of treatments? In another double-blind randomized trial of treatments for heart failure, the effect of enalapril was compared with that of a placebo [11]. Of 1285 patients randomly assigned to receive the ACE inhibitor, 48% (613) died or were admitted to hospital for worsening heart failure, whereas 57% (736/1284) of patients who received placebo died or required hospital care. The point estimate of the difference in death or hospital admission for heart failure was 10%, and the 95% confidence interval was 6% to 14%. Thus, the smallest true effect of the ACE inhibitor that is compatible with the data is a 6% (or about 1 in 17) reduction in the number of patients with these adverse outcomes. If it is considered worth while to treat 17 patients in order to prevent one death or heart failure, this trial is definitive. If, before using a drug, you require a reduction of more than 6% in the proportion of patients who are spared death or heart failure, a larger trial (with a correspondingly narrower confidence interval) would be required.

WAS THE TRIAL LARGE ENOUGH?

Confidence intervals provide a way of answering the question Was the trial large enough? We illustrate this approach in Fig. 1. Each of the distribution curves represents the results of one hypothetical randomized trial of an experimental treatment to reduce mortality (trials A, B, C and D). The vertical line at 0% represents a risk reduction of 0: a result at this value means that mortality in the experimental and control groups is exactly the same. Values to the right of the vertical line represent results in which the experimental group had a lower mortality than the control group; to the left of the vertical line, results in which the experimental group fared worse, with a higher mortality than the control group.

The highest point of each distribution represents the result actually observed (the point estimate). In trials A and B the investigators observed that mortality was 5% lower in the experimental group than in the control group. In trials C and D they observed that mortality was 1% higher in the experimental group than in the

control group.

The distributions of the likelihood of possible true results of each trial are based on the point estimate and the size of the sample. The point estimate is the single value that is most likely to represent the true effect. As you can see, values farther from the results observed are less likely than values closer to the point estimate to represent the true difference in mortality.

Now, suppose we assume that an absolute reduction in mortality of greater than 1% means that treatment is warranted (that is, such a result is clinically important), and a reduction of less than 1% means that treatment is not warranted (that is, the result is trivial). For example, if the experimental treatment results in a true reduction in mortality from 5% to 4% or less, we would want to use the treatment. If, on the other hand, the true reduction in mortality was from 5% to 4.5%, we would consider the benefit of the experimental treatment not to be worth the associated toxic effects and cost. What are the implications of this decision for the interpretation of the results of the four studies?

In trial A the entire distribution and, hence, the entire 95% confidence interval lies above the threshold risk reduction of 1%. We can therefore be confident that the true treatment effect is above our threshold, and we have a definitive "positive" trial. That is, we can be very confident that the true reduction in risk is greater -- probably appreciably greater -- than 1%; this leaves little doubt that we should administer the treatment to our patients. The sample size in this trial was adequate to show that the treatment provides a clinically important benefit.

Trial B has the same point estimate of treatment effect as trial A (5%) and is also "positive" ($p < 0.05$). In a hypothesis test, the null hypothesis would be rejected. However, more than 2.5% of the distribution is to the left of the 1% threshold. In other words, the 95% confidence interval includes values less than 1%. This means that the data are consistent with an absolute risk reduction of less than 1%, so we are left with some doubt that the treatment effect is really greater than our threshold. This trial is still "positive," but its results are not definitive. The sample in this trial was inadequate to establish definitively the appropriateness of administering the experimental treatment.

Trial C is "negative"; its results would not lead to the rejection of the null hypothesis in a hypothesis test. The investigators observed mortality 1% higher in the treatment than in the control group. The entire distribution and, therefore, the 95% confidence interval lie to the left of our 1% threshold. Because the upper limit of the distribution is 1%, we can be very confident that, if there is a positive effect, it is trivial. The trial has excluded any clinically important benefit of treatment, and it can be considered definitive. We can therefore decide against the use of the experimental treatment, at least for this type of patient.

The result of trial D shows the same difference in absolute risk as that of trial C: mortality 1% higher in the experimental than in the control group. However, trial D had a smaller sample and, as a result, a much wider distribution and confidence interval. Since an appreciable portion of the confidence interval lies to the right of our 1% threshold, it is plausible (although unlikely) that the true effect of the experimental treatment is a reduction in mortality of greater than 1%. Although we would refrain from using this treatment (indeed, the most likely conclusion is that it kills people), we cannot completely dismiss it. Trial D was not definitive, and a trial involving more patients is required to exclude a clinically important treatment effect.

CONCLUSIONS

We can restate our interpretation of confidence intervals as follows. In a "positive" trial -- one that establishes that the effect of treatment is greater than zero -- look at the lower boundary of the confidence interval to determine whether the size of the sample is adequate. The lower boundary represents the smallest plausible treatment effect compatible with the data. If it is greater than the smallest difference that is clinically important, the sample size is adequate and the trial definitive. However, if it is less than this smallest important difference, the trial is not definitive and further trials are required. In a "negative" trial -- the results of which do not exclude the possibility that the treatment has no effect -- look at the upper boundary of the confidence interval to determine whether the size of the sample is adequate. If the upper boundary -- the largest treatment effect compatible with the data -- is less than the smallest difference that is clinically important, the size of the sample is adequate, and the trial is definitively negative. If the upper boundary exceeds the smallest difference considered important, there may be an important positive treatment effect, the trial is not definitive, and further trials are required.

In this discussion we have examined absolute differences in proportions of patients who died while receiving two different treatments. In the next article in this series, we will explain how to interpret other ways investigators present treatment effects, including odds ratios and relative risk.

REFERENCES

1. Simon R: Confidence intervals for reporting results of clinical trials. Ann Intern Med 1986; 105: 429-435 [Medline Link] [BIOSIS Previews Link] [Context Link]

2. Gardner MJ, Altman DG (eds): Statistics with Confidence: Confidence Intervals and Statistical Guidelines, British Medical Journal, London, England, 1989 [Context Link]

3. Bulpitt CJ: Confidence intervals. Lancet 1987; 1: 494-497 [Medline Link] [BIOSIS Previews Link] [Context Link]

4. Pocock SJ, Hughes MD: Estimation issues in clinical trials and overviews. Stat Med 1990; 9: 657-671 [Medline Link] [BIOSIS Previews Link] [Context Link]

5. Braitman LE: Confidence intervals assess both clinical significance and statistical significance. Ann Intern Med 1991; 114: 515-517 [Medline Link] [BIOSIS Previews Link] [Context Link]

6. Cohn JN, Johnson G, Ziesche S et al: A comparison of enalapril with hydralazine-isosorbide dinitrate in the treatment of chronic congestive heart failure. N Engl J Med 1991; 325: 303-310 [Medline Link] [BIOSIS Previews Link] [Context Link]

7. Oxman AD, Sackett DL, Guyatt GH and the Evidence-Based Medicine Working Group: Users' guides to the medical literature: I. How to get started. JAMA 1993; 270: 2093-2095 [Context Link]

8. Guyatt GH, Sackett DL, Cook DJ and the Evidence-Based Working Group: Users' guides to the medical literature: II. How to use an article about therapy or prevention. A. Are the results of the study valid? JAMA 1993; 270: 2598-2601 [Context Link]

9. Guyatt GH, Sackett DL, Cook DJ and the Evidence-Based Working Group: Users' guides to the medical literature: II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? JAMA 1994; 271: 59-63 [Context Link]

10. Mulrow CD, Mulrow JP, Linn WD et al: Relative efficacy of vasodilator therapy in chronic congestive heart failure. JAMA 1988; 259: 3422-3426 [Medline Link] [BIOSIS Previews Link] [Context Link]

11. The SOLVD Investigators: Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. N Engl J Med 1991; 325: 293-302 [Context Link]

12. Sackett DL, Haynes RB, Guyatt GH et al: Clinical Epidemiology, a Basic Science for Clinical Medicine, Little, Brown and Company, Boston, 1991: 218-220 [Context Link]

13. Britton M, Helmers C, Samuelsson K: High-dose salicylic acid after cerebral infarction: a Swedish co-operative study. Stroke 1987; 18: 325 [BIOSIS Previews Link] [Context Link]

14. Oxman AD, Cook DJ, Guyatt GH and the Evidence-Based Medicine Working Group: Users' guides to the medical literature: VI. How to use an overview. JAMA 1994; 272: 1367-1371 [Context Link]

15. Antiplatelet trialists' collaboration: Secondary prevention of vascular disease by prolonged antiplatelet treatment. BMJ 1988; 296: 320-331 [Context Link]

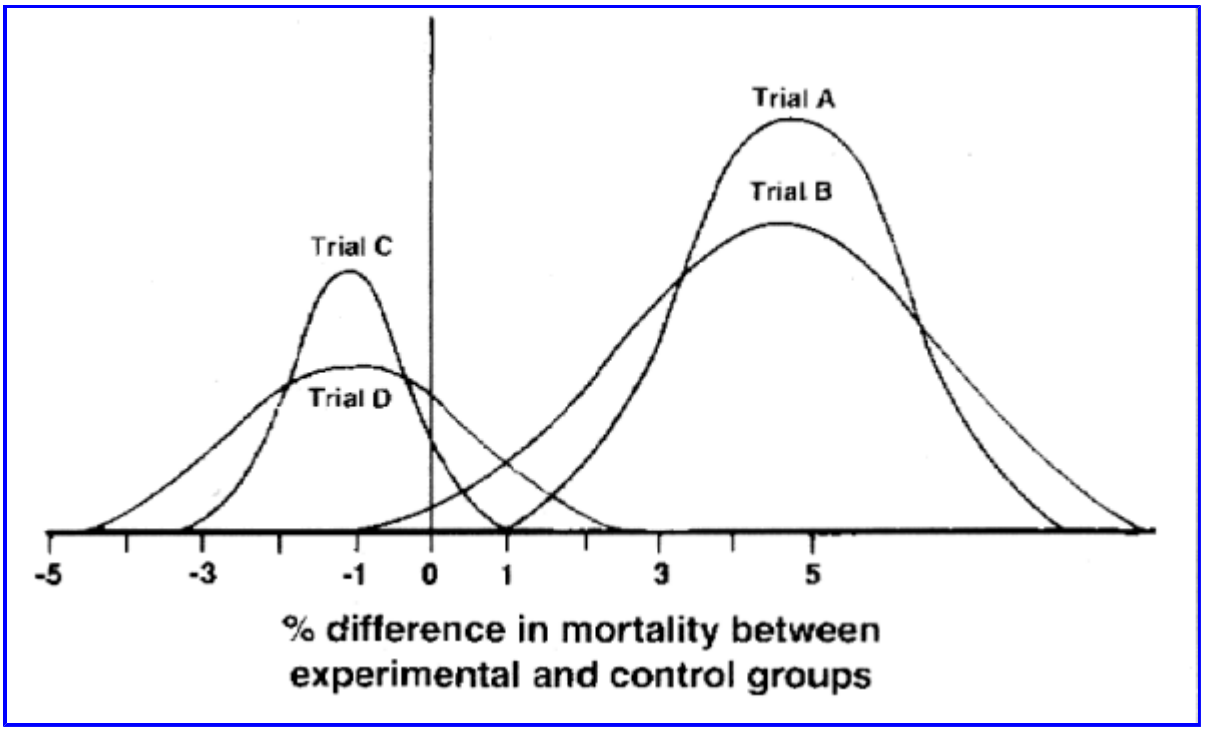

Figure 1. Distributions of the likehood of the true results of four trials (A, B, C and D). Trial A is a definitive and trial B a nondefinitive positive trial. Trial C is a definitive and trial D a nondefinitive negative trial

Accession Number: 00002792-199501150-00017

Volume 152(3) 1 February 1995 pp 351-357

Basic Statistics for Clinicians: 3. Assessing the Effects of Treatment: Measures of Association

[Statistics]

Jaeschke, Roman; Guyatt, Gordon; Shannon, Harry; Walter, Stephen; Cook, Deborah; Heddle, Nancy

From the departments of Clinical Epidemiology and Biostatistics, Medicine and Pathology, McMaster University, Hamilton, Ont.

Dr. Cook is a recipient of a Career Scientist Award from the Ontario Ministry of Health. Dr. Walter is the recipient of a National Health Scientist Award from Health Canada.

Reprint requests to: Dr. Gordon Guyatt, Rm. 2C12, McMaster University Health Sciences Centre, 1200 Main St. W, Hamilton ON L8N 3Z5.

This is the third article in a series of four, to appear in the January and February 1995 issues of CMAJ.

Outline

Graphics

Abstract

In the third of a series of four articles the authors show the calculation of measures of association and discuss their usefulness in clinical decision making.From the rates of death or other "events" in experimental and control groups in a clinical trial, we can calculate the relative risk (RR) of the event after the experimental treatment, expressed as a percentage of the risk without such treatment. The absolute risk reduction (ARR) is the difference in the risk of an event between the groups. The relative risk reduction is the percentage of the baseline risk (the risk of an event in the control patients) removed as a result of therapy. The odds ratio (OR), which is the measure of choice in case-control studies, gives the ratio of the odds of an event in the experimental group to those in the control group. The OR and the RR provide limited information in reporting the results of prospective trials because they do not reflect changes in the baseline risk. The ARR and the number needed to treat, which tells the clinician how many patients need to be treated to prevent one event, reflect both the baseline risk and the relative risk reduction. If the timing of events is important -- to determine whether treatment extends life, for example -- survival curves are used to show when events occur over time.

The reader familiar with the first two articles in this series will, when presented with the results of a clinical trial, know how to discover the range within which the treatment effect likely lies. This treatment effect is worth considering if it comes from a study that is valid [1]. In this article, we explore the ways investigators and representatives of pharmaceutical companies may present the results of a trial.

When clinicians look at the results of clinical trials they are interested in the association between a treatment and an outcome. There may be no association; for example, there may be no difference in mean values of an indicator -- such as blood pressure -- between groups, or the same risk of an adverse event -- such as death -- in both groups. Alternatively, the trial results may show a decreased risk of adverse outcomes in patients receiving the experimental treatment. In a study examining a putatively harmful agent there may be no increase in risk among patients in a group exposed to the agent in comparison with those in a control group or an association between exposure and an adverse event, which suggests that the agent is indeed harmful. In this article, we examine how one can express the magnitude of these associations.

When investigators present results that show a difference in the mean value of a clinical measurement between two groups, the interpretation is usually straightforward. However, when they present results that show the proportion of

patients who suffered an adverse event in each group, interpretation may be more difficult. In this situation they may express the strength of the association as a relative risk, an absolute risk reduction or an odds ratio. Understanding these measures is challenging and important; they will provide the focus of this article. We will examine the relative merits of the different measures of association and show how they can lead clinicians to different conclusions.

INTRODUCING THE 2 x 2 TABLE

A crucial concept in analysing the efficacy of therapeutic interventions is the "event." Analysis often examines the proportion of patients who suffered a particular outcome (the "event") in the treatment and control groups. This is always true when the outcome is clearly a dichotomous variable -- that is, a discrete event that either occurs or does not occur. Examples of dichotomous outcomes are the occurrence of negative events, such as stroke, myocardial infarction, death or recurrence of cancer, or positive events, such as ulcer healing or resolution of symptoms. Not only an event's occurrence but also its timing may be important. We will return to this issue later.

Even if the results are not of a yes-or-no form, investigators sometimes choose to present them as if they were. Investigators may present variables such as duration of exercise before chest pain develops, number of episodes of angina per month, change in lung function or number of visits to the emergency room as mean values in each of the two groups. However, they may also transform these values into dichotomous data by specifying a threshold or degree of change that constitutes an important improvement or deterioration and then examining the proportion of patients above and below this threshold. For example, investigators in one study used forced expiratory volume in 1 second (FEV_1) to assess the efficacy of therapy with corticosteroids taken orally by patients with a chronic stable airflow limitation; they defined an "event" as an improvement in FEV_1 of more than 20% over the baseline value [2].

The results of trials with dichotomous outcomes can usually be presented in a form of 2 x 2 table Table 1. For instance, in a randomized trial investigators compared rates of death among patients with bleeding esophageal varices controlled by either endoscopic ligation or sclerotherapy [3]. After a mean follow-up period of 10 months, 18 of 64 patients assigned to ligation died, as did 29 of 65 patients assigned to sclerotherapy. Table 2summarizes the data from this trial in a 2 x 2 table.

	Outcome	
Exposure	Yes	No
Yes	A	B
No	C	D

Table 1. Sample 2 x 2 table

	Outcome, no. of patients		Total no. of patients treated
Intervention	Death	Survival	
Ligation	18	46	64
Sclerotherapy	29	36	65

*Reprinted with permission from *N Engl J Med* 1992; 326: 1527–1532.

Table 2. Results from a randomized trial comparing treatment of bleeding esphageal varices with endoscopic sclerotherapy and with ligation

RELATIVE RISK

The first thing we can determine from the 2 x 2 Table is that the risk of an event (death, in this case) was 28.1% (18/64) in the ligation group and 44.6% (29/65) in the sclerotherapy group. The ratio of these risks is called the relative risk (RR) or the risk ratio. This value tells us the risk of the event after the experimental treatment (in this case, ligation), as a percentage of the original risk (in this case, the risk of death after sclerotherapy). From Table 1, the formula for calculating the RR from the data gathered is (A/(A + B))/(C/(C + D)). In our example, the RR of death after receiving initial ligation compared with sclerotherapy is 18/64 (the risk in the ligation group) divided by 29/65 (the risk in the sclerotherapy group), which equals 63%. That is, the risk of death after ligation is about two thirds as great as the risk of death after sclerotherapy.

ABSOLUTE RISK REDUCTION

The difference in the risk of the outcome between patients who have undergone one therapy and those who have undergone another is called the absolute or attributable risk reduction (ARR) or the risk difference. The formula for its calculation, from Table 1, is (C/(C + D)) - (A/(A + B)). This measure tells us the percentage of patients who are spared the adverse outcome as a result of having received the experimental rather than the control therapy. In our example, the ARR is 0.446 minus 0.281, which equals 0.165, or 16.5%.

RELATIVE RISK REDUCTION

Another measure used to assess the effectiveness of treatment is relative risk reduction (RRR). One considers first the risk of an adverse event among patients taking the placebo or, if two therapies are being compared, the risk among patients receiving the standard or inferior therapy. This is called the baseline risk. The relative risk reduction is an estimate of the percentage of baseline risk that is removed as a result of the therapy; it is calculated as the ARR between the treatment and control groups, divided by the absolute risk among patients in the control group; from Table 1, ((C/(C + D)) - (A/(A + B)))/(C/(C + C)). In our example, the RRR is calculated by dividing 16.5% (the ARR) by 44.6% (the risk among patients receiving sclero-therapy), which equals 37%. One may also derive the RRR by subtracting the RR from 1. In our example, the RRR is equal to 1 minus 0.63, or 0.37 (37%).

ODDS RATIO

Instead of looking at the risk of an event, we could estimate the odds of an event occurring. In our example, the odds of death after ligation are 18 (death) versus 46 (survival), or 18/46 (A/B), and the odds of death after sclero-therapy are 29 versus 36 (C/D). The formula for the ratio of these odds -- called, not surprisingly, the odds ratio (OR) -- is (A/C)/(B/D). In our example, this calculation yields (18/46)/(29/36), which equals 0.49.

The OR is probably less familiar to physicians than risk or RR. However, the OR is usually the measure of choice in the analysis of case-control studies. In general, the OR has certain optimal statistical properties that make it the fundamental measure of association in many types of studies [4]. These statistical advantages may be particularly important when data from several studies are combined, as they are in a meta-analysis. Among such advantages, the comparison of risk represented by the OR does not depend on whether the investigator chose to determine the risk of an event occurring (e.g., death) or not occurring (e.g., survival). This is not true for relative risk. In some situations the OR and the RR will be close -- for example, in case-control studies of a rare disease.

RR VERSUS OR VERSUS ARR: WHY THE FUSS?

The important distinction among the ARR, the RR and the OR may be illustrated by modifying the death rates in each of the two treatment groups shown in Table 2. In the explanation that follows, the reader should note that the effect on the various expressions of risk depends on the way the death rates are changed. We could alter the death rates by the same absolute amount in each group, by the same relative amount, or in some other way.

There is some evidence that, when treatment reduces the rate of death, the reduction in rates or proportion of deaths will often be similar in each subgroup of patients [5,6]. In our example, if we assume that the number of patients who died decreased by 50% in both groups, the risk of death in the ligation group would decrease from 28% to 14% and in the sclerotherapy group from 44.6% to 22.3%. The RR would be 14/22.3 or 0.63 -- the same as before. The OR would be (9/55)/(14.5/51) or 0.58, which differs moderately from the OR based on the higher death rate (0.49), and is closer to the RR. The ARR would decrease from 16.5% to approximately 8%. Thus, a decrease in the proportion of patients who died in both groups by a factor of two leaves the RR unchanged, results in a moderate increase in the OR and reduces the ARR by a factor of two. This example highlights the fact that the same RR can be associated with very different ORs and ARRs. A major change in the risk of an adverse event without treatment (or, as in this case, with the inferior treatment) will not be reflected in the RR or the OR; in contrast, the ARR changes markedly with a change in the baseline risk.

Hence, the RR and the OR do not tell us the magnitude of the absolute risk. An RR of 33% may mean that the treatment reduces the risk of an adverse outcome from 3% to 1% or from 60% to 20%. The clinical implications of these risk reductions are very different. Consider a therapy with severe side effects. If such side effects occur in 5% of patients treated, and the treatment reduces the probability of an adverse outcome from 3% to 1%, we probably will not institute this therapy. However, we may be willing to accept this incidence of side effects if the therapy reduces the probability of an adverse outcome from 60% to 20%. In the latter situation, of every 100 patients treated 40 would benefit and 5 would suffer side effects -- a trade-off that most would consider worth while.

The RRR behaves the same way as the RR: it does not reflect the change in the underlying risk in the control population. In our example, if the incidence of adverse events decreased by approximately 50% in both groups, the RRR would be the same as it was at the previous incidence rate: (22.3 - 14)/22.3 or 0.37. The RRR therefore shares with the RR the disadvantage of not reflecting the baseline risk.

These observations depend on the assumption that the death rates in the two groups change by the same proportion. If these changes are not proportional the

conclusions may be different. For instance, suppose that the rates of death between the two groups differ by 10 percentage points; for example, if the death rates are 80% and 90%, respectively, the RR is 0.8/0.9 or 89%, the RRR 11%, the ARR 10% and the OR 0.44. If the rates of death then decrease by 50 percentage points in each group, to 30% and 40% respectively, the RR would be 0.3/0.4 or 75%, the RRR 25%, the ARR 10% and the OR 0.64. In this case, the ARR remains constant and thus does not reflect the change in the magnitude of risk without therapy. In contrast, the other indices differ in the two cases and hence reflect the change in the baseline risk.

NUMBER NEEDED TO TREAT

The number needed to treat (NNT) is the most recently introduced measure of treatment efficacy [7]. Let us return to our 2 x 2 tables for a short exercise. In Table 2we see that the risk of death in the ligation group is 28.1% and in the sclerotherapy group 44.6%. Therefore, treating 100 patients with ligation rather than sclerotherapy will save the lives of between 15 and 16 patients, as shown by the ARR. If treating 100 patients prevents 16 adverse events, how many patients do we need to treat to prevent 1 event? The answer is 100 divided by 16, which yields approximately 6. This is the NNT. One can also arrive at this number by taking the reciprocal of the ARR (1/ARR). Since the NNT is related to the ARR, it is not surprising that the NNT also changes with a change in the underlying risk.

The NNT is directly related to the proportion of patients in the control group who suffer an adverse event. For instance, if the incidence of these events (the baseline risk) decreased by a factor of two and the RRR remained constant, treating 100 patients with ligation would mean that 8 events had been avoided, and the NNT would double, from 6 to 12. In general, the NNT changes inversely in relation to the baseline risk. If the risk of an adverse event doubles, we need treat only half as many patients to prevent the same number of adverse events; if the risk decreases by a factor of four, we must treat four times as many patients to achieve the same result.

BACK TO THE 2 x 2 TABLE

The data we have presented so far could have been derived from the original 2 x 2 table Table 2. The ARR and its reciprocal, the NNT, incorporate the influence of any change in baseline risk, but they do not tell us the magnitude of the baseline risk. For example, an ARR of 5% (and a corresponding NNT of 20) may represent reduction of the risk of death from 10% to 5% or from 50% to 45%. The RR and RRR do not take into account the baseline risk, and the clinical utility of these measures suffers as a result.

Whichever way we choose to express the efficacy of a treatment, we must keep

in mind that the 2 x 2 Table reflectsresults at a given time. Therefore, our comments on the RR, the ARR, the RRR, the OR and the NNT must be qualified by giving them a time frame. For example, we must say that use of ligation rather than sclerotherapy for a mean period of 10 months resulted in an ARR of 17% and an NNT of 6. The results could be different if the duration of observation was very short, in which case there was little time for an event such as death to occur, or very long, in which case it is much more likely that an event will occur (e.g., if the outcome is death, after 100 years of follow-up all of the patients will have died).

CONFIDENCE INTERVALS

We have presented all of the measures of association for treatment with ligation versus sclerotherapy as if they represented the true effect. As we pointed out in the previous article in this series, the results of any experiment are an estimate of the truth. The true effect of treatment may actually be greater or less than what we observed. The confidence interval tells us, within the bounds of plausibility, how much greater or smaller the true effect is likely to be. Confidence intervals can be calculated for each of the measures of association we have discussed.

SURVIVAL DATA

As we pointed out, the analysis of a 2 x 2 Table is an examination of the data at a specific time. Such analysis is satisfactory if we are investigating events that occur within relatively short periods and if all patients are followed for the same duration. However, in longer-term studies we are interested not only in the number of events but also in their timing. We may, for instance, wish to know whether therapy for a fatal condition such as severe congestive heart failure or unresectable lung cancer delays death.

When the timing of events is important, the results can be presented in several 2 x 2 tables constructed at certain points after the beginning of the study. In this sense, Table 2showed the situation after a mean of 10 months of follow-up. Similar tables could be constructed to show the fate of all patients at given times after their enrolment in the trial, i.e., at 1 week, 1 month, 3 months or whatever intervals we choose. An analysis of accumulated data that takes into account the timing of events is called survival analysis. Despite the name, such analysis is not restricted to deaths; any discrete event may be studied in this way.

The survival curve of a group of patients shows the status of the patients at different times after a defined starting point [8]. In Fig. 1, we show an example of a survival curve taken from a trial of treatments of bleeding varices. Although the mean follow-up period in this trial was 286 days, the survival curve extends beyond this time, presumably to a point at which the number of patients still at risk is sufficient to make reasonably confident predictions. At a later point, prediction

would become very imprecise because there would be too few patients to estimate the probability of survival. This imprecision can be captured by confidence intervals or bands extending above and below the survival curves.

Hypothesis tests can be applied to survival curves, the null hypothesis being that there is no difference between two curves. In the first article in this series, we described how an analysis based on hypothesis testing can be adjusted or corrected for differences in the two groups at the baseline. If one group were older (and thus had a higher risk of the adverse outcome) or had less severe disease (and thus had a lower risk), the investigators could conduct an analysis that takes into account these differences. Such an analysis tells us, in effect, what would have happened if the two groups had comparable risks of adverse outcomes at the start of the trial.

CASE-CONTROL STUDIES

The examples we have used so far have been prospective randomized controlled trials. In such trials we start with an experimental group of patients who are subject to an intervention and a control group of patients who are not. The investigators follow the patients over time and record the incidence of events. The process is similar in prospective cohort studies, although in this study design the "exposure" or treatment is not controlled by the investigators. Instead of being assigned to receive or not receive the intervention, patients are chosen, sampled or classified according to whether they were or were not exposed to the treatment or risk factor. In both randomized trials and prospective cohort studies we can calculate risks, ARRs and RRs.

In case-control studies participants are chosen or sampled not according to whether they have been exposed to the treatment or risk factor but on the basis of whether they have experienced an event. Participants start the study with or without the event rather than with or without the exposure or intervention. Patients with the adverse outcome -- be it stroke, myocardial infarction or cancer -- are compared with control patients who have not suffered the outcome. The investigators wish to determine if any factor seems to be more common in one of these groups than in the other.

In one case-control study investigators examined whether the use of sun-beds or sun-lamps increased the risk of melanoma [9]. They identified 583 patients with melanoma and 608 control patients. The control and case patients had similar distributions of age, sex and region of residence. The results for men and women were presented separately (those for men are shown in Table 3.

Ever exposed to sun-beds or sun-lamps	No. of patients	
	Case	Control
Yes	67	41
No	210	242

*Reproduced with permission from Walter SD, Marrett LD, From L et al: The association of cutaneous malignant melanoma with the use of sunbeds and sunlamps. *Am J Epidemiol* 199; 131: 232–243.

Table 3. Results from a case-control study of the association between meanoma and the use of sunbeds and sun-lamps

If the information in Table 3came from a prospective cohort study or randomized controlled trial we could begin by calculating the risk of an event in the experimental and control groups. However, this would not make sense in a case-control study because the number of patients who did not have melanoma was chosen by the investigators. For calculation of the RR we need to know the population at risk, and this information is not available in a case-control study.

The only measure of association that makes sense in a case-control study is the OR. One can investigate whether the odds of having been exposed to sun-beds or sun-lamps among the patients with melanoma are the same as the odds of exposure among the control patients. In the study the odds were 67/210 in the patients with melanoma and 41/242 in the control patients. The odds ratio is therefore (67/210)/(41/242) or 1.88 (95% confidence interval (CI) 1.20 to 2.98), which suggests an association between the use of sun-beds or sun-lamps and melanoma. The fact that the CI does not include 1.0 means that the association is unlikely to be due to chance.

Even if the association were not due to chance, this does not necessarily mean that the sun-beds or sun-lamps were the cause of melanoma in these patients. Potential explanations could include higher recollection of use of these devices among patients with melanoma (recall bias), longer exposure to sun among these patients or different skin colour. (In fact, in this study the investigators addressed many of these possible explanations.) Confirmatory studies would be needed to be confident that exposure to sun-beds or sun-lamps was the cause of melanoma.

WHICH MEASURE OF ASSOCIATION IS BEST?

In randomized trials and cohort studies, investigators can usually choose from several measures of association. Which should the reader hope to see? We believe that the best option is to show all of the data, in the form of 2 x 2 tables or life tables (deaths or other events during follow-up presented in tabular form), and then consider both the relative and absolute figures. As the reader examines the results, she or he will find the ARR and its reciprocal, the NNT, the most useful measures for deciding whether to institute treatment. As we have discussed, the RR and the RRR do not take baseline risk into account and can therefore be misleading.

In fact, clinicians make different decisions depending on the way the results are reported. Clinicians consistently judge a therapy to be less effective when the results are presented in the form of the NNT than when any other measure of association is used [10]--13.

INTERPRETING STUDY RESULTS

We complete this exposition by reviewing the results of a landmark study -- the Lipid Research Clinics Coronary Primary Prevention Trial -- of the usefulness of therapy to lower serum cholesterol levels [14]. In this randomized, placebo-controlled trial the investigators tested the hypothesis that a reduction in cholesterol levels reduces the incidence of coronary heart disease (CHD). They followed 3806 asymptomatic middle-aged men with primary hyper-cholesterolemia (serum cholesterol levels above the 95th percentile), of whom one third were smokers, for a mean period of 7.4 years. Patients in one group received cholestyramine (24 g/d) and those in the other a placebo. The main outcome measures (events) were death due to CHD and nonfatal myocardial infarction. After 7.4 years of follow-up the results showed an ARR of 1.71% (95% CI -0.11% to 3.53%) and an NNT of 58 (the 95% CI for the NNT would include the fact that the therapy causes one death in 935 treated patients and requires treatment of 28 patients to save one life). The original report did not provide CIs for the RR and the ARR. We used the original data to calculate these measures and the associated CIs, so our point estimates differ slightly from the adjusted estimates given in the original report.

The risk of an event was 9.8% among the patients taking a placebo and 8.1% among those receiving cholestyramine. The RR of an event for those taking cholestyramine versus those taking a placebo was 83% (95% CI 68% to 101%). The use of cholestyramine was associated with a 17% reduction in the incidence of an event (RRR), with a 95% CI from a 33% reduction in risk to a 1% increase in risk, and with prevention of 17 primary events per 1000 patients treated. Therefore, 58 patients (100/1.7) needed to be treated for 7 years to prevent one primary event.

In addition to calculating the NNT, one could also consider resources expended to prevent an event. The cost of a month's supply of cholestyramine is $120.49.

The cost of the drug required to prevent one event is 58 (the NNT) x 7 years of follow-up x 12 months per year x $120.49 for a 1-month supply = $587 027.28. Alternatively, to prevent one event, patients need to take 24 g/d x 58 (NNT) x 365 days per year x 7 years of follow-up = 3 556 560 g, approximately 3.56 tonnes to swallow of cholestyramine.

If one considered only patients with a lower risk of CHD (younger men, women, nonsmokers and those with cholesterol levels that are elevated but not in the top 95th percentile) the NNT would rise. It is not surprising that advertisements promoting the use of cholesterol-lowering drugs cite the RRR rather than the ARR or the NNT and do not mention the cost per event prevented.

The results of this study provide another caution for the clinician. The results we have described are based on the incidence of both fatal and nonfatal coronary events. However, the death rates shown in this study were similar in the two groups: there were 71 deaths among patients receiving placebo and 68 among patients receiving cholestyramine. Furthermore, when investigators have examined all trials of drug therapy for lowering cholesterol, they have found a possible association between administration of these agents and death from causes other than cardiovascular disease [15]. As this result highlights, the wary user of the medical literature must be sure that all relevant outcomes are reported [16].

ARRs are easy to calculate, as is their reciprocal, the NNT. If the NNT is not presented in trial results, clinicians who wish to get the best sense of the effect of an intervention should take the trouble to determine the number of patients they need to treat to prevent an event as well as the cost and toxic effects associated with treatment of that number of patients. These measures will help clinicians to weigh the benefits and costs of treatments.

REFERENCES

1. Guyatt GH, Sackett DL, Cook DJ and the Evidence-based Medicine Working Group: Users' guides to the medical literature: II. How to use an article about therapy or prevention. A. Are the results of the study valid? JAMA 1993; 270: 2598-2601 [Context Link]

2. Callahan CM, Dittus RS, Katz BP: Oral corticosteroid therapy for patients with stable chronic obstructive pulmonary disease. A meta-analysis. Ann Intern Med 1991; 114: 216-223 [Medline Link] [BIOSIS Previews Link] [Context Link]

3. Stiegman GV, Goff JS, Michaletz-Onody PA et al: Endoscopic sclerotherapy as compared with endoscopic ligation for bleeding esophageal varices. N Engl J Med 1992; 326: 1527-1532 [Medline Link] [BIOSIS Previews Link] [Context Link]

4. Laird NM, Mosteller F: Some statistical methods for combining experimental results. Int J Technol Assess Health Care 1990; 6: 5-30 [Medline Link] [Context Link]

5. Oxman AD, Guyatt GH: A consumer's guide to subgroup analysis. Ann Intern Med 1992; 116: 78-84

[Medline Link] [BIOSIS Previews Link] [Context Link]

6. Yusuf S, Wittes J, Probstfield J et al: Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA 1991; 266: 93-98 [Medline Link] [BIOSIS Previews Link] [Context Link]

7. Laupacis A, Sackett DL, Roberts RS: An assessment of clinically useful measures of the consequences of treatment. N Engl J Med 1988; 318: 1728-1733. [Medline Link] [BIOSIS Previews Link] [Context Link]

8. Coldman AJ, Elwood JM: Examining survival data. Can Med Assoc J 1979; 121: 1065-1071 [Medline Link] [BIOSIS Previews Link] [Context Link]

9. Walter SD, Marrett LD, From L et al: The association of cutaneous malignant melanoma with the use of sunbeds and sunlamps. Am J Epidemiol 1990; 131: 232-243 [Medline Link] [Context Link]

10. Forrow L, Taylor WC, Arnold RM: Absolutely relative: How research results are summarized can affect treatment decisions. Am J Med 1992; 92: 121-124 [Medline Link] [BIOSIS Previews Link] [Context Link]

11. Naylor CD, Chen E, Strauss B: Measured enthusiasm: Does the method of reporting trial results alter perceptions of therapeutic effectiveness? Ann Intern Med 1992; 117: 916-921 [Medline Link] [BIOSIS Previews Link]

12. Bucher HC, Weinbacher M, Gyr K: Influence of method of

13. reporting study results on decision of physicians to prescribe

14. drugs to lower cholesterol concentration. BMJ 1994; 309: 761-764 [Context Link]

15. Bobbio M, Demichells B, Glustetto C: Completeness of reporting trial results: effect on physicians' willingness to prescribe. Lancet 1994; 343: 1209-1211 [Medline Link] [Context Link]

16. The Lipid Research Clinics Coronary Primary Prevention Trial results. I. Reduction in incidence of coronary heart disease. JAMA 1984; 251: 351-364 [Medline Link] [BIOSIS Previews Link] [Context Link]

17. Davey Smith G, Pekkanen J: Should there be a moratorium on the use of cholesterol lowering drugs? BMJ 1992; 304: 431-434 [Medline Link] [BIOSIS Previews Link]

18. Guyatt GH, Sackett DL, Cook DJ and the Evidence-Based Medicine Working Group: Users' guides to the medical literature: II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? JAMA 1994: 271: 59-63

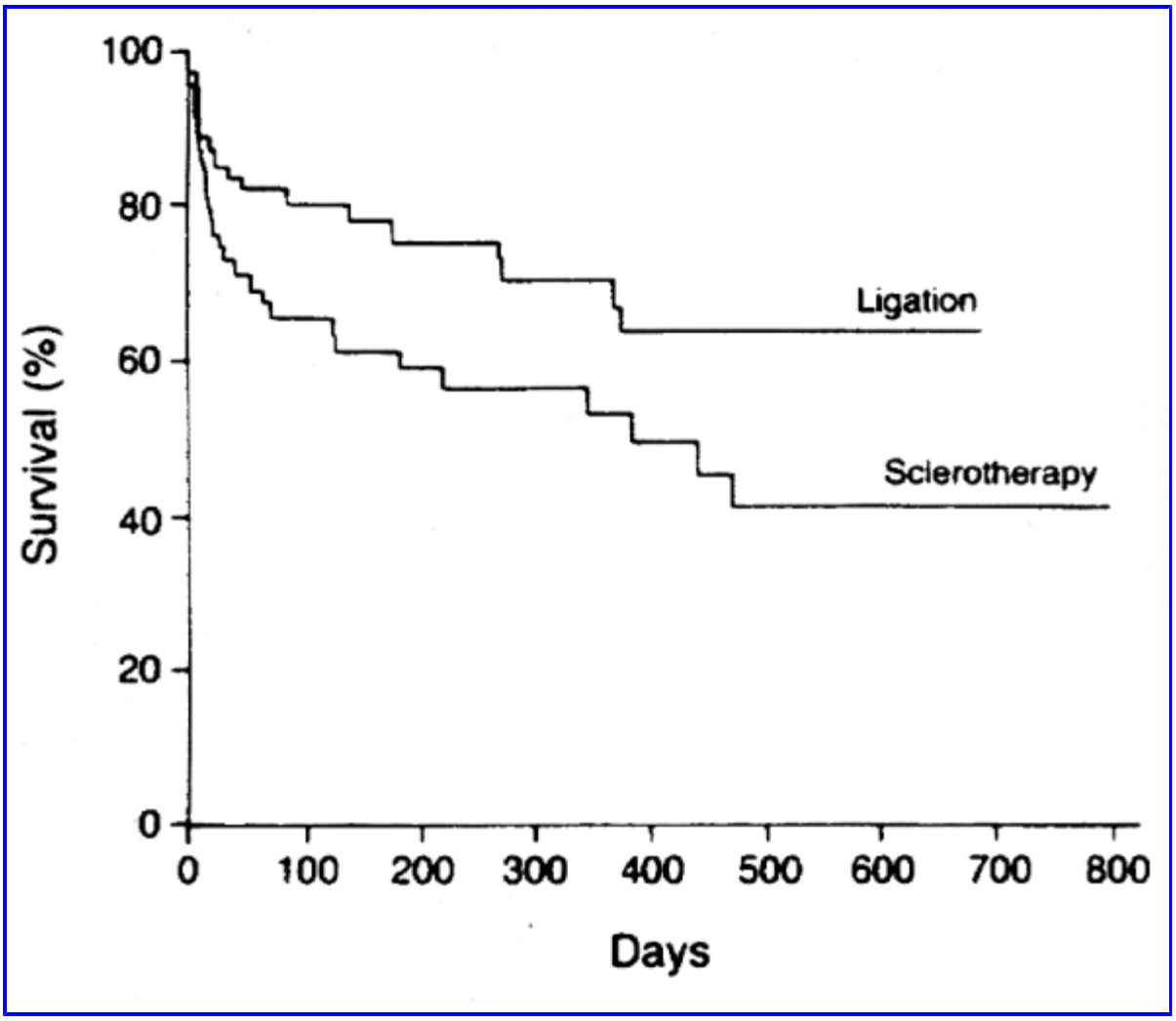

Figure 1. Survival curves showing percentages of patients who survived after treatment of bleeding esophageal varices with ligation and with sclerotherapy. Reprinted with permission from N Engl J Med 1992; 326: 1527-1532

Accession Number: 00002792-199502010-00024

Volume 152(4) 15 February 1995 pp 497-504

Basic Statistics for Clinicians: 4. Correlation and Regression

[Statistics]

Guyatt, Gordon; Walter, Stephen; Shannon, Harry; Cook, Deborah; Jaeschke, Roman; Heddle, Nancy

From the departments of Clinical Epidemiology and Biostatistics, Medicine and Pathology, McMaster University, Hamilton, Ont.

Dr. Cook is a recipient of a Career Scientist Award from the Ontario Ministry of Health. Dr. Walter is the recipient of a National Health Scientist Award from Health Canada.

Reprint requests to: Dr. Gordon Guyatt, Rm. 2C12, McMaster University Health Sciences Centre, 1200 Main St. W, Hamilton ON L8N 3Z5.

This is the final article in a series of four that began in the Jan. 1, 1995, issue of CMAJ.

Outline

Graphics

Abstract

Correlation and regression help us to understand the relation between variables and to predict patients' status in regard to a particular variable of interest.Correlation examines the strength of the relation between two variables, neither of which is considered the variable one is trying to predict (the target

variable). Regression analysis examines the ability of one or more factors, called independent variables, to predict a patient's status in regard to the target or dependent variable. Independent and dependent variables may be continuous (taking a wide range of values) or binary (dichotomous, yielding yes-or-no results). Regression models can be used to construct clinical prediction rules that help to guide clinical decisions. In considering regression and correlation, clinicians should pay more attention to the magnitude of the correlation or the predictive power of the regression than to whether the relation is statistically significant.

Clinicians are sometimes interested in the relation between different factors or "variables." How well does a relative's impression of a patient's symptoms and well-being predict the patient's own report? How strong is the relation between a patient's physical well-being and emotional function? In answering these questions, our goal is to enhance our understanding and consider the implications for action. If the relation between patients' perceptions and those of patients' relatives is not a strong one, the clinician must obtain both perspectives on a situation. If physical and emotional function are only weakly related, then clinicians must probe both areas thoroughly.

Clinicians may be even more interested in making predictions or causal inferences than in understanding the relation between phenomena. Which clinical features of patients with chest pain presenting to the emergency department predict whether they have a myocardial infarction? What determines how dyspneic we feel when we exercise or when we suffer from a cardiac or respiratory illness? Can we predict which critically ill patients will tolerate weaning from a ventilator and which will not?

We refer to the first issue -- understanding the magnitude of the relation between different variables or phenomena -- as "correlation." We call the statistical techniques for exploring the second issue -- making a prediction or causal inference -- "regression." In this final article in our series we will provide illustrations of the use of correlation and regression in medical literature.

CORRELATION

Traditionally, we measure the exercise capacity of patients with cardiac and respiratory illnesses with the use of a treadmill or cycle ergometer. About 20 years ago, investigators interested in respiratory disease began to use a simpler test that is more closely related to day-to-day activity [1]. In this test, patients are asked to cover as much ground as they can in a specified time (typically 6 minutes) walking in an enclosed corridor. There are several reasons why we may be interested in the strength of the relation between the 6-minute walk test and conventional laboratory

measures of exercise capacity. If the results of these tests are strongly related, we could substitute one test for the other. In addition, the strength of the relation could tell us how well exercise capacity, determined by laboratory measures, predicts patients' ability to undertake physically demanding activities of daily living.

What do we mean by the strength of the relation between two variables? A relation is strong when patients who obtain high scores on the first variable also obtain high scores on the second, those who have intermediate scores on the first variable also show intermediate values on the second, and those who have low scores on one measure score low on the other. By contrast, if patients who have low scores on one measure are equally likely to have high or low scores on another, the relation between the two variables is poor or weak.

We can gain a sense of the strength of the correlation by examining a graph that relates patients' scores on the two measures. Fig. 1 presents a scatterplot of the results of the walk test and of the cycle ergometer exercise test. The data for this graph, and for the subsequent analyses involving walk-test results, are taken from three studies of patients with chronic airflow limitation [2,3,4]. Each point on the scatterplot is for an individual patient and presents two pieces of information: the patient's walk-test score and cycle ergometer exercise score. The walk-test results are continuous; however, the cycle ergometer results tend to take only certain values because patients usually stop the test at the end of a particular level. From Fig. 1, one can see that, in general, patients who have a high score on the walk test tend to have a high score on the cycle ergometer exercise test, and patients who have a low score on the cycle ergometer test tend to have a low score on the walk test as well. Yet one can find patients who are exceptions, scoring higher than most other patients on one test and not as high on the other.

These data represent a moderately strong relation between two variables, the walk test and the cycle ergometer exercise test. The strength of the relation can be summarized in a single number, the correlation coefficient (r). The correlation coefficient can range from -1.0 (the strongest possible negative relation -- the patient with the highest score on one test has the lowest score on the other) to 1.0 (the strongest possible positive relation). A correlation coefficient of 0 denotes no relation at all between the two variables: patients with a high score on one test have the same range of scores on the other test as those with a low score on the first test. The scatterplot of data with a correlation coefficient of 0 looks like a starry sky (without the constellations).

The correlation coefficient assumes a straight-line relation between the variables. However, there may be a relation between the variables that does not take the form of a straight line. For example, values of the variables may rise together, but one may rise more slowly than the other for low values and more

quickly than the other for high values. If there is a strong relation, but it is not a straight line, the correlation coefficient may be misleading. In our example, the relation does appear to approximate a straight line, and the value of r for the correlation between the walk test and the cycle ergometer test is 0.5.

This value for r indicates a moderately strong correlation, but is it strong enough? It depends on how we wish to apply the information. If we were thinking of substituting the walk test for the cycle ergometer test (after all, the walk test is much simpler to carry out) we would be disappointed. A correlation of 0.8 or higher is required for us to feel comfortable with that kind of substitution. If the correlation is any lower than 0.8, there is too great a risk that a patient with a high score on the walk test would have mediocre or low score on the cycle ergometer test or vice versa. However, if we assume that the walk test provides a good indication of exercise capacity in day-to-day life, the moderately strong correlation suggests that the result of the cycle ergometer test also tells us something, although not as much, about day-to-day exercise capacity.

You will often see a p value provided with a correlation coefficient (the first article in this series discusses the interpretation of p values). This p value is determined from a hypothesis test, with the null hypothesis being that the true correlation between the two measures is 0. Thus, the p value represents the probability that, if the true correlation were 0, a relation as strong as or stronger than the one we actually observed would have occurred by chance. The smaller the p value, the less likely it is that chance explains the apparent relation between the two measures.

The p value depends not only on the strength of the relation but also on the sample size. In this case, we had data on the results of the walk test and the cycle ergometer test from 179 patients and a correlation coefficient of 0.5, which yields a p value of less than 0.0001. A relation can be very weak, but if the sample is large enough the p value may be small. For instance, with a sample of 500, we reach the conventional threshold for statistical significance ($p = 0.05$) when the correlation coefficient is only 0.10.

In a previous article in this series we pointed out that, in evaluating treatment effects, the size of the effect and the confidence interval tend to be much more informative than p values. The same is true of correlations: the magnitude of the correlation and the confidence interval are the key values. The 95% confidence interval for the correlation between the results of the walk test and of the laboratory exercise test is 0.38 to 0.60.

REGRESSION

As clinicians, we are often interested in prediction: we wish to know which

patient will get a disease (such as coronary artery disease) and which will not, and which patient will fare well (returning home after a hip fracture rather than remaining in an institution) and which will fare poorly. Regression analysis is useful in addressing these sorts of issues. We will once again use the walk test to illustrate the concepts involved in statistical regression.

Predicting walk-test scores

Let us consider an investigation in which the goal is to predict patients' walk-test scores from more easily measured variables: sex, height and a measure of lung function (forced expiratory volume in 1 second (FEV ($_1$))). Alternatively, we can think of the investigation as an examination of a causal hypothesis. To what extent are patients' walk-test scores determined by their sex, height and lung function? Either way, we have a target or response variable that we call the dependent variable (in this case the walk-test score) because it is influenced or determined by other variables or factors. We also have the explanatory or predictor variables, which we call independent variables -- sex, height and FEV_1.

Fig. 2, a histogram of the walk-test scores for 219 patients with long-term lung disease, shows that these scores vary widely. If we had to predict an individual patient's walk-test score without any other information, our best guess would be the mean score for all patients (394 m). For many patients, however, this prediction would be well off the mark.

Fig. 3 shows the relation between FEV_1 and walk-test scores. There is a relation between the two variables, although it is not as strong as that between the walk-test score and the exercise-test score, examined earlier (Fig. 1). Thus, some of the variation in walk-test scores seems to be explained by, or attributable to, the patient's FEV_1. We can construct an Equation thatpredicts the walk-test score as a function of FEV ($_1$). Because there is only one independent variable, we call this a univariate or simple regression [5].

In regression equations we generally refer to the predictor variable as x and the target variable as y. The Equation assumesa straight-line fit between the FEV_1 and the walk-test score, and specifies the point at which the straight line meets the y-axis (the intercept) and the steepness of the line (the slope). In this case, the regression Equation isy = 298 + 108x, where y is the walk-test score in metres, 298 is the intercept, 108 is the slope of the line and x is the FEV ($_1$) in litres. In this case, the intercept of 298 has little practical meaning: it predicts the walk-test score of a patient with an FEV_1 of 0 L. The slope of 108 does, however, have meaning: it predicts that, for every increase in FEV_1 of 1 L, the patient will walk 108 m farther. The regression line corresponding to this Equation isshown in Fig. 3.

We can now examine the correlation between the two variables, and whether it can be explained by chance. The correlation coefficient is 0.4, and, since p is 0.0001, chance is a very unlikely explanation for this relation. Thus, we conclude that FEV_1 explains or accounts for a statistically significant proportion of the variation in walk-test scores.

We can also examine the relation between the walk-test score and the patients' sex (Fig. 4). Although there is considerable variation in scores among men and among women, men tend to have higher scores than women. If we had to predict a man's score, we would choose the mean score for the men (410 m), and we would choose the mean score for the women (363 m) to predict a woman's score.

Is the apparent relation between sex and the walk-test score due to chance? One way of answering this question is to construct a simple regression Equation withthe walk-test score as the dependent variable and the sex of the patient as the independent variable. As it turns out, chance is an unlikely explanation of the relation between sex and the walk-test score ($p = 0.0005$).

As these examples show, the independent variable in a regression Equation canbe an either/or variable, such as sex (male or female), which we call a dichotomous variable, or a variable that can theoretically take any value, such as FEV_1, which we call a continuous variable.

In Fig. 5 we have divided the men from the women, and for each sex we have separated the patients into groups with a high FEV_1 and a low FEV_1. Although there is still a range of scores within each of these four groups, the range is narrower. When we use the mean of any group as our best guess for the walk-test score of any member of that group, we will be closer to the true value than if we had used the mean for all patients.

Fig. 5 illustrates how we can take into account more than one independent variable in explaining or predicting the dependent variable. We can construct a mathematical model that explains or predicts the walk-test score by simultaneously considering all of the independent variables; this is called a multivariate or multiple regression equation.

We can learn several things from such an equation. First, we can determine whether the independent variables from the univariate equations each make independent contributions to explaining the variation. In this example, we consider first the independent variable with the strongest relation to the dependent variable, then the variable with the next strongest relation and so on. FEV_1 and sex make independent contributions to explaining walk test ($p < 0.0001$ for FEV_1 and $p = 0.03$ for sex in the multiple regression analysis), but height (which was significant

at the p = 0.02 level when considered in a univariate regression) does not.

If we had chosen the FEV_1 and the peak expiratory flow rate as independent variables, they would both have shown significant associations with walk-test score. However, the FEV_1 and the peak expiratory flow rate are very strongly associated with one another; therefore, they are unlikely to provide independent contributions to explaining the variation in walk-test scores. In other words, once we take the FEV_1 into account, the peak flow rates are not likely to be of any help in predicting walk-test scores; likewise, if we first took the peak flow rate into account, the FEV_1 would not provide further explanatory power in our model. Similarly, height was a significant predictor of walk-test score when considered alone, but it was no longer significant in the multiple regression because of its correlation with sex and FEV_1.

We have emphasized that the p value associated with a correlation provides little information about the strength of the relation between two values; the correlation coefficient is required. Similarly, the knowledge that sex and FEV_1 independently explain some of the variation in walk-test scores tells us little about the power of our predictive model. We can get some sense of the model's predictive power from Fig. 5. Although the distributions of walk-test scores in the four subgroups differ appreciably, there is considerable overlap. The regression Equation cantell us the proportion of the variation in the dependent variable (that is, the differences in walk-test scores among patients) associated with each of the independent variables (sex and FEV_1) and, therefore, the proportion explained by the entire model. In this case, the FEV_1 explains 15% of the variation when it is the first variable entered into the model, sex explains an additional 2% of the variation once the FEV_1 is in the model already, and the overall model explains 17% of the variation. We can therefore conclude that many other factors we have not measured (and perhaps cannot measure) determine how far people with long-term lung disease can walk in 6 minutes. Other regression analyses have found that patients' experience of the intensity of their exertion as well as their perception of the severity of their illness may be more powerful determinants of walk-test distance than their FEV_1 [6].

In this example, the dependent variable -- the walk-test score -- was continuous. Because this regression analysis assumes a straight-line fit between the independent and dependent variable, and the dependent variable is continuous, we refer to the analysis as "linear regression." In our next example, the dependent variable is dichotomous. Investigators sometimes use the term "logistic regression" to refer to such models because they are based on logarithmic equations.

Predicting clinically important gastrointestinal bleeding

We have recently considered whether we could predict which critically ill patients were at risk of clinically important gastrointestinal bleeding [7]. In this example, the dependent variable was whether patients had had a clinically important bleed. When the dependent variable is dichotomous we use a logistic regression. The independent variables included whether patients were breathing independently or required ventilator support and the presence or absence of coagulopathy, sepsis, hypotension, hepatic failure and renal failure.

In the study we followed 2252 critically ill patients and determined which of them had clinically important gastrointestinal bleeding. Table 1, which contains some of the results, shows that in univariate logistic regression analyses many of the independent variables were significantly associated with clinically important bleeding. For several variables, the odds ratio (discussed in a previous article in this series), which indicates the strength of the association, was large. However, when we constructed a multiple logistic regression equation, only two of the independent variables -- ventilator support and coagulopathy -- were significantly and independently associated with bleeding. All of the other variables that had predicted bleeding in the univariate analysis were correlated with either ventilation or coagulopathy and were not statistically significant in the multiple regression analysis. Of the patients who were not supported by a ventilator, 0.2% (3/1597) had an episode of clinically significant bleeding, whereas 4.6% (30/655) of those being supported by a ventilator had such an episode. Of those with no coagulopathy 0.6% (10/1792) had an episode of bleeding, whereas of those with coagulopathy 5.1% (23/455) had such an episode.

Risk factors	Odds ratio (*p* value)	
	Simple regression analysis	Multiple regression analysis*
Respiratory failure	25.5 (< 0.0001)	15.6 (< 0.0001)
Coagulopathy	9.5 (< 0.0001)	4.3 (0.0002)
Hypotension	5.0 (0.03)	2.1 (0.08)
Sepsis	7.3 (< 0.0001)	NS
Hepatic failure	6.5 (< 0.0001)	NS
Renal failure	4.6 (< 0.0001)	NS
Enteral feeding	3.8 (0.0002)	NS
Administration of steroids	3.7 (0.0004)	NS
Transplantation of an organ	3.6 (0.006)	NS
Therapy with anticoagulants	3.3 (0.004)	NS

*NS = not significant.

Table 1. Odds ratios and p values for risk factors for clinically important gastrointestinal bleeding in critically ill patients, calculated with use of simple and multiple logistic regression analysis

Our main clinical interest was identification of a subgroup with a risk of bleeding low enough that prophylaxis could be withheld. In an analysis separate from the regression analysis, but suggested by its results, we divided the patients into two groups, those who were neither supported by a ventilator nor had coagulopathy, in whom the incidence of bleeding was only 0.14% (2/1405), and those who were either supported by a ventilator or had coagulopathy, of whom 3.7% (31/847) had an episode of bleeding. Prophylaxis may reasonably be withheld from patients in the former group.

CONCLUSION

Correlation examines the strength of the relation between two variables, neither of which is necessarily considered the target variable. Regression examines the strength of the relation between one or more predictor variables and a target variable. Regression can be very useful in formulating predictive models such as the risk of myocardial infarction in patients presenting with chest pain, [8] the risk of cardiac events in patients undergoing noncardiac surgery, [9] or the risk of gastrointestinal bleeding in critically ill patients. Such predictive models can help us make clinical decisions. Whether you are considering a correlation between variables or a regression analysis, you should consider not only the statistical

significance of the relation but also its magnitude or strength, in terms of the proportion of variation explained by the model or the extent to which groups with very different risks can be specified.

We thank Derek King, BMath, for conducting the original analyses reported in this article and for preparing the figures.

REFERENCES

1. McGavin CR, Gupta SP, McHardy GJR: Twelve-minute walking test for assessing disability in chronic bronchitis. BMJ 1976; 1: 822-823 [Medline Link] [Context Link]

2. Guyatt GH, Berman LB, Townsend M: Long-term outcome after respiratory rehabilitation. Can Med Assoc J 1987; 137: 1089-1095 [Medline Link] [CINAHL Link] [BIOSIS Previews Link] [Context Link]

3. Guyatt GH, Keller J, Singer J et al: A controlled trial of respiratory muscle training in chronic airflow limitation. Thorax 1992; 47: 598-602 [Medline Link] [BIOSIS Previews Link] [Context Link]

4. Goldstein RS, Gort EH, Stubbing D et al: Randomized controlled trial of respiratory rehabilitation. Lancet 1994; 344: 1394-1397 [Medline Link] [CINAHL Link] [Context Link]

5. Godfrey K: Simple linear regression. N Engl J Med 1985; 313: 1629-1636 [Medline Link] [BIOSIS Previews Link] [Context Link]

6. Morgan AD, Peck DF, Buchanan DR et al: Effect of attitudes and beliefs on exercise tolerance in chronic bronchitis. BMJ 1983; 286: 171-173 [BIOSIS Previews Link] [Context Link]

7. Cook DJ, Fuller HD, Guyatt GH et al: Risk factors for gastrointestinal bleeding in critically ill patients. N Engl J Med 1994; 330: 377-381 **[Fulltext Link]** [Medline Link] [CINAHL Link] [BIOSIS Previews Link] [Context Link]

8. Pozen MW, D'Agostino RB, Selker HP et al: A predictive instrument to improve coronary-care-unit admission practices in acute ischemic heart disease. N Engl J Med 1984; 310: 1273-1288 [Medline Link] [BIOSIS Previews Link] [Context Link]

9. Detsky AS, Abrams HB, McLaughlin JR et al: Predicting cardiac complications in patients undergoing non-cardiac surgery. J Gen Intern Med 1986; 1: 211-219 [Medline Link] [BIOSIS Previews Link] [Context Link]

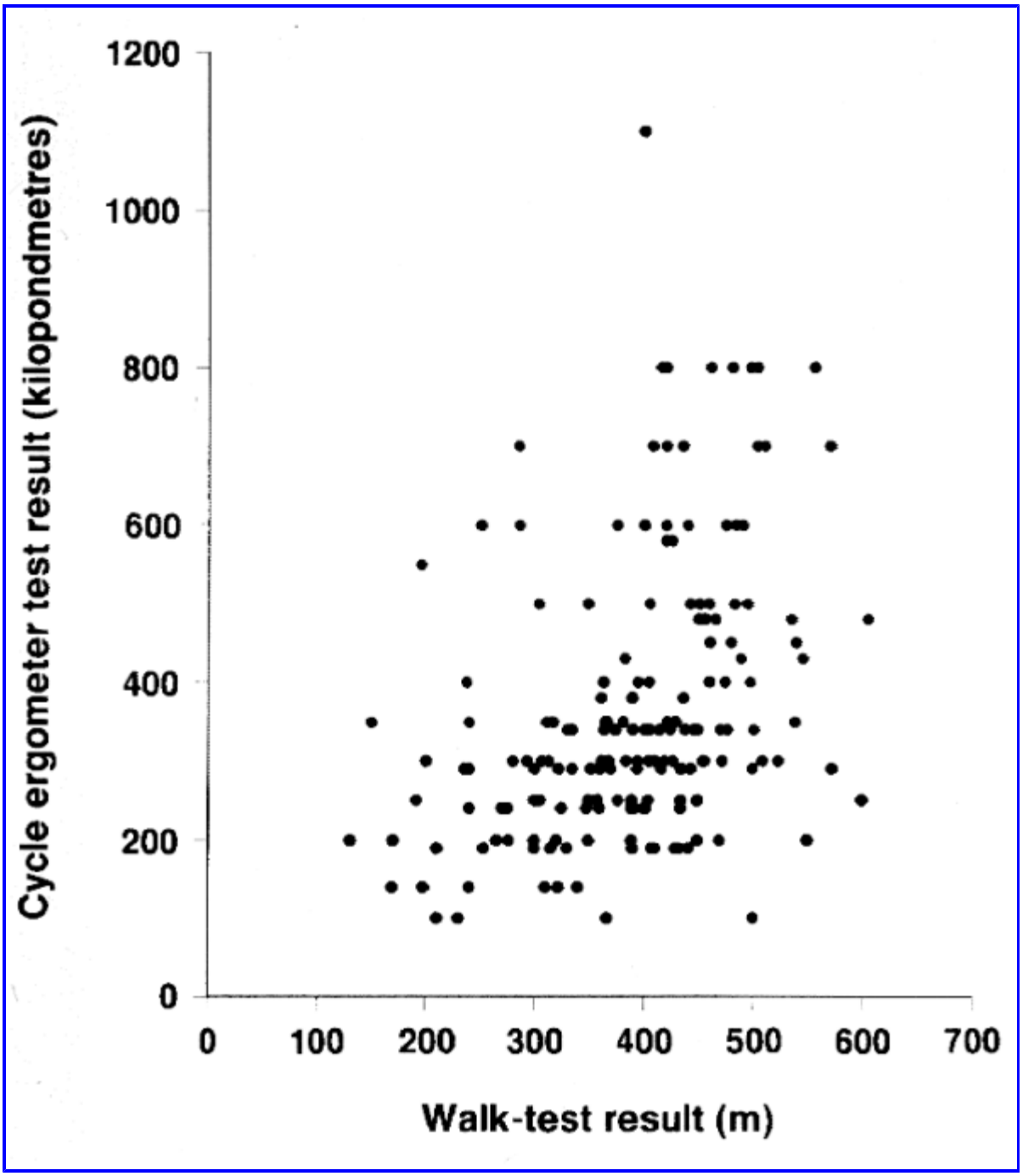

Figure 1. Scatterplot of the results of the 6-minute walk test and the cycle ergometer exercise test for 179 patients. Each point gives the results for one patient

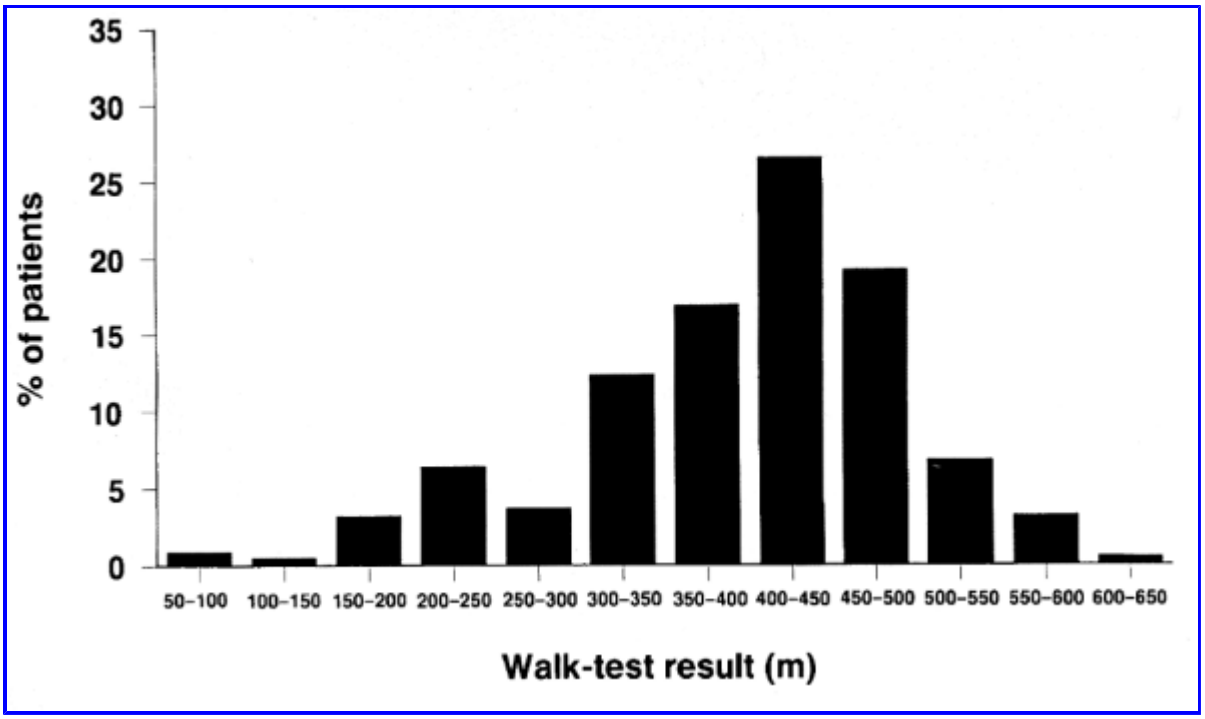

Figure 2. Distribution of 6-minute walk-test results in a sample of 219 patients

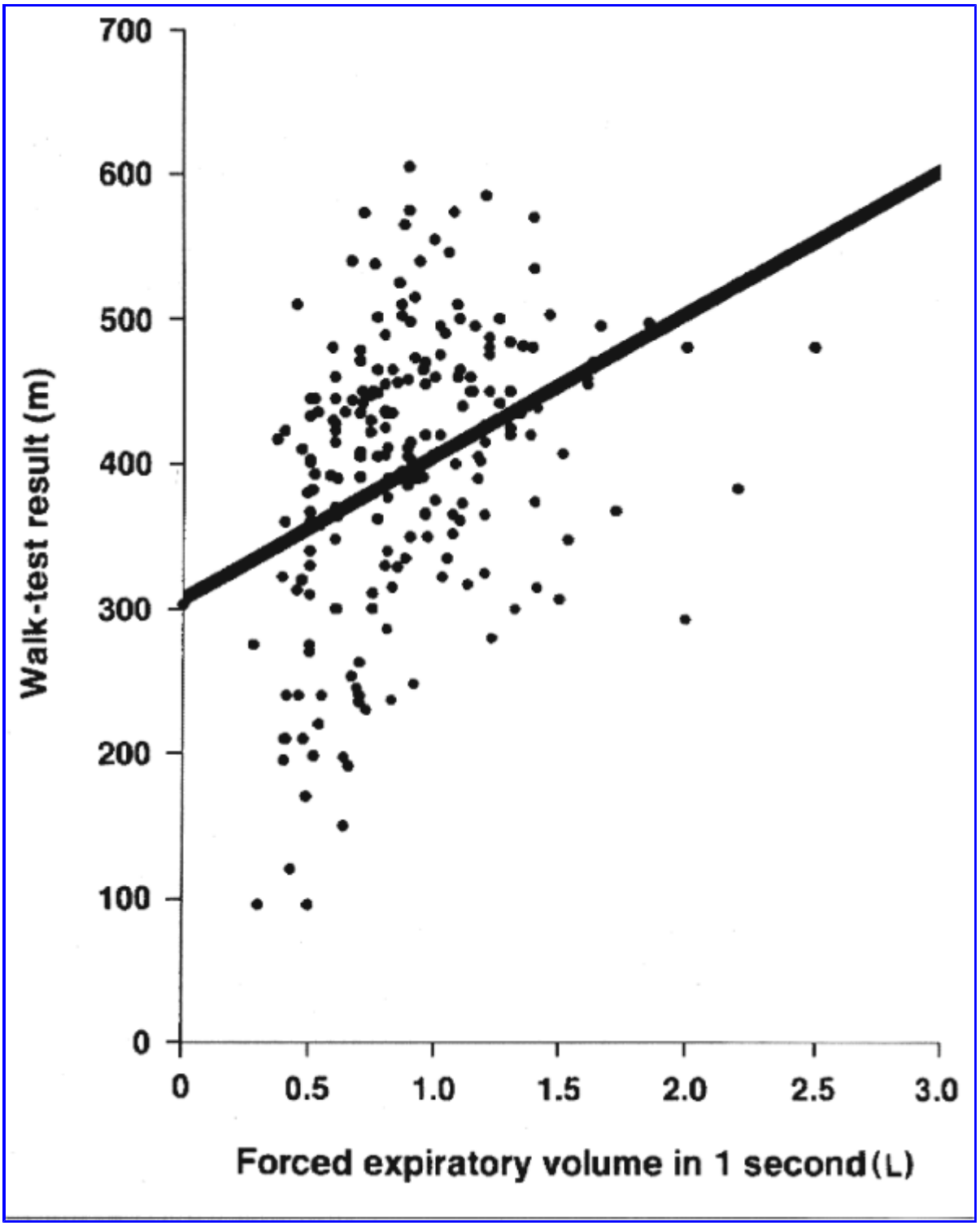

Figure 3. Scatterplot for the expiratory volume in 1 second and of the 6-mminute walk-test results for 219 patients. Each point gives the results for one patient

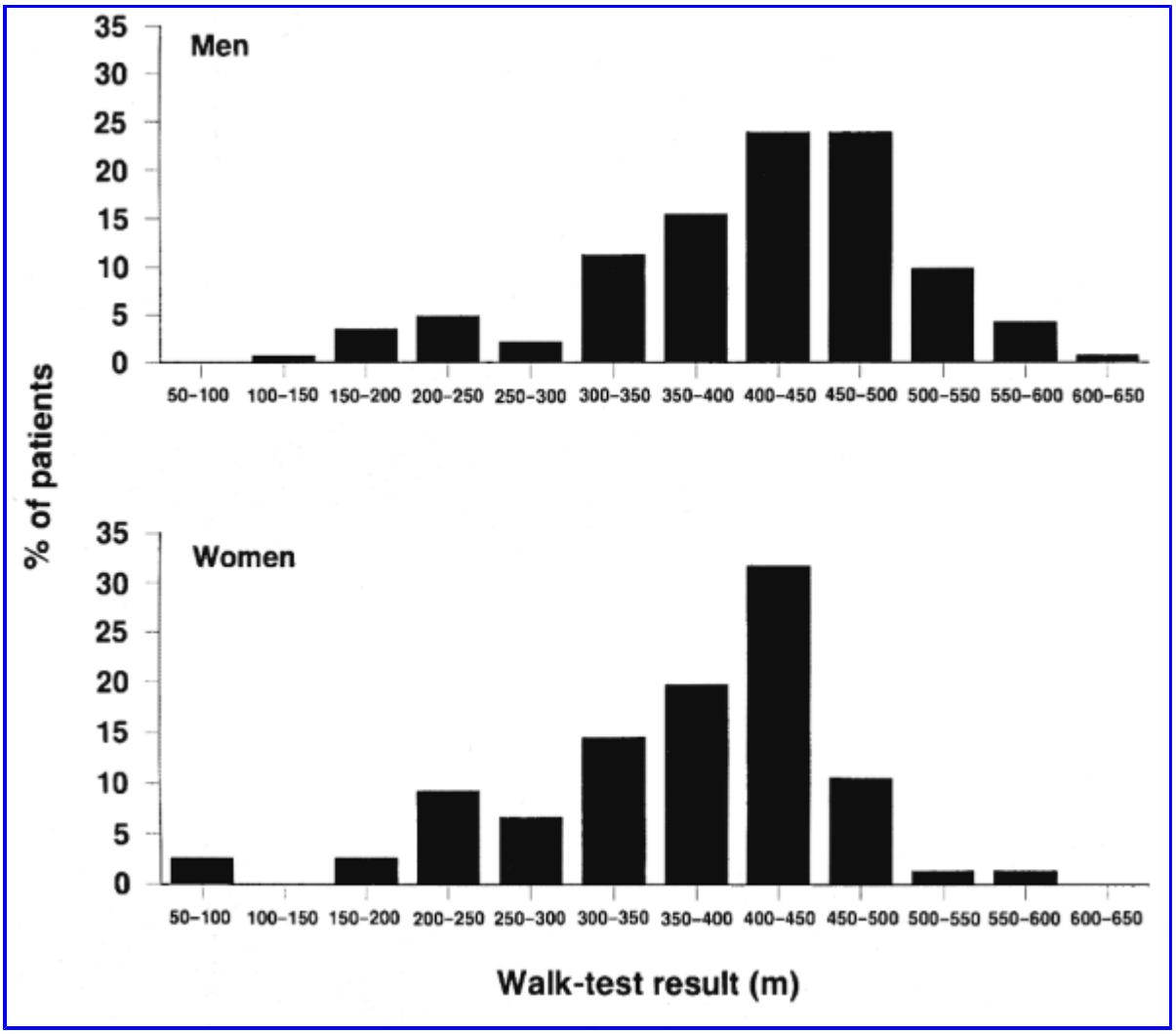

Figure 4. Distribution of the 6-minute walk-test results in men (top) and women (bottom) from the sample of 219 patients

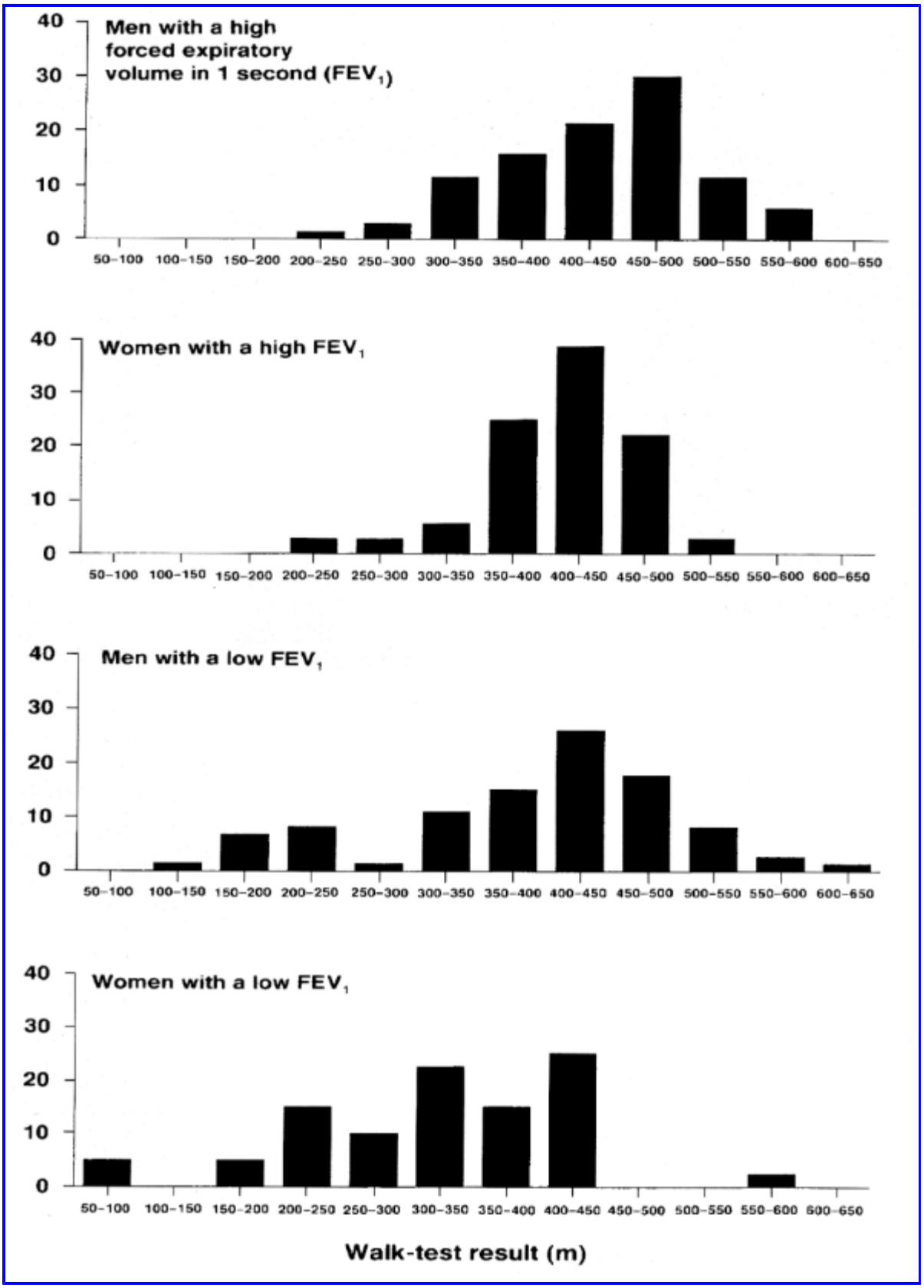

Figure 5. Distribution of the 6-minute walk test results in men and women with high FEV, (top), and women with low FEV, (bottom) from the sample of 219 patients

Accession Number: 00002792-199502150-00020